

# ESTIMATING EQUATION METHODS FOR LONGITUDINAL AND SURVIVAL DATA

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

David Young Clement

January 2011

© 2011 David Young Clement  
ALL RIGHTS RESERVED

# ESTIMATING EQUATION METHODS FOR LONGITUDINAL AND SURVIVAL DATA

David Young Clement, Ph.D.

Cornell University 2011

This thesis analyzes censored data in recurrent event, longitudinal, and survival settings. In Chapter 2, a straightforward, flexible methodology is proposed to estimate parameters indexing the conditional means and variances of the inter-event times in a recurrent event process. In Chapter 3, we analyze discretely and informatively observed multivariate continuous longitudinal data; missingness and terminal events are introduced in Chapter 4. In Chapters 3 and 4, the inter-event times are considered a nuisance and the goal is to estimate parameters driving the longitudinal process. To do this, we propose an innovative conditional estimating equation that can model individual trajectories. Finally, Chapter 5 uses these subject-specific trajectories to estimate parameters indexing the terminal event process and predict future survival for arbitrary subjects.

## BIOGRAPHICAL SKETCH

The author was born in 1981 in Toronto, Ontario and from an early age began following the Toronto Blue Jays baseball team. Even before he was old enough to stay up and watch their games, he would read (and memorize) the statistics in the morning newspaper: a career working with numbers, even if they weren't baseball-related, was never far from his mind.

He matriculated at the University of Toronto in 2000 without knowing exactly what field to ultimately pursue: mathematics, computer science, or chemistry. But after taking a couple of undergraduate statistics courses, he realized this was the best field for him - more mathematical than computer science and chemistry but not as excruciatingly technical as pure mathematics can sometimes be, and also with real world applications that would allow him to study problems from all aspects of society while still enjoying some computer programming. Upon obtaining his B.Sc. in Statistics and Applied Mathematics in 2004, he entered the Statistical Science MS/Ph.D. program at Cornell University.

This document is dedicated to my parents, Christine and Maurice Clement.

## ACKNOWLEDGEMENTS

First and foremost I would like to thank my advisor, Dr. Robert Strawderman, who was always available to answer questions regarding the material contained herein, particularly the material in Chapter 2, which was written when my research skills were still developing. His encyclopedic knowledge of the whole field of statistics, especially survival analysis, and his constant nudging towards stronger methodological content helped make this thesis more compelling than it would have been had I been left to my own devices. I feel I have matured as a statistician largely from his example, and the research I produce in the future will be stronger having learned from him.

I would also like to thank my other committee members, Dr. Giles Hooker and Dr. Martin Wells, for helpful comments, especially after my A Exam. They were both willing to spend time and give advice whenever I needed it. I acknowledge Dr. Haiqun Lin, a fellow Cornell graduate, for providing the HUD-VASH data from her paper, Dr. George Kaysen for providing the longitudinal protein data, Dr. Lei Liu for providing the medical cost data, and Dr. Mark Cowen for not only providing the cardiac data, but quickly answering many questions about the medicine involved, which greatly enhanced my understanding of the dataset and its collection.

Finally, the partial support of National Institutes of Health grant R01 GM056182 is gratefully acknowledged.

## TABLE OF CONTENTS

Biographical Sketch . . . . .	iii
Dedication . . . . .	iv
Acknowledgements . . . . .	v
Table of Contents . . . . .	vi
List of Tables . . . . .	ix
List of Figures . . . . .	xi
<b>1 Introduction</b>	<b>1</b>
<b>2 Conditional GEE for recurrent event gap times</b>	<b>5</b>
2.1 Introduction . . . . .	5
2.2 Methodology . . . . .	7
2.2.1 Notation and model . . . . .	7
2.2.2 Estimation with “full” and observed data: single subject . . . . .	9
2.2.3 Estimation and inference with observed data: $n$ subjects . . . . .	13
2.2.4 Connections to Murphy et al. (1995) . . . . .	14
2.2.5 Connections with GEE . . . . .	16
2.3 Simulations . . . . .	18
2.4 Data Analysis: Recurrent asthma in children . . . . .	23
2.5 Discussion . . . . .	30
<b>3 Marginal and conditional estimating equations for multivariate longitudinal data subject to censoring</b>	<b>33</b>
3.1 Introduction . . . . .	33
3.2 Methodology . . . . .	37
3.2.1 Notation . . . . .	37
3.2.2 Marginal model . . . . .	38
3.2.3 Conditional model . . . . .	43
3.3 Simulations . . . . .	46
3.4 Data Analyses . . . . .	50
3.4.1 HUD-VASH . . . . .	50
3.4.2 Multivariate longitudinal protein . . . . .	53
3.5 Discussion . . . . .	58
<b>4 Conditional estimating equations for multivariate longitudinal data subject to censoring, failure, and missingness</b>	<b>61</b>
4.1 Introduction . . . . .	61
4.2 Methodology . . . . .	63
4.2.1 Introducing failure times . . . . .	63
4.2.2 Introducing missingness . . . . .	64
4.2.3 CEE with failure times and missingness . . . . .	66
4.2.4 IPSW digression . . . . .	68

4.3	Simulations . . . . .	69
4.4	Data Analysis: Medical costs . . . . .	72
4.4.1	CHF data introduction . . . . .	72
4.4.2	Previous models considered by Liu and coauthors for the CHF data . . . . .	74
4.4.3	A new model for the CHF data . . . . .	74
4.5	Discussion . . . . .	78
<b>5</b>	<b>Survival prediction based on discretely observed covariates with miss- ingness</b>	<b>81</b>
5.1	Introduction . . . . .	81
5.2	Covariate distribution at arbitrary times . . . . .	84
5.3	Methodology . . . . .	85
5.3.1	Notation and assumptions . . . . .	85
5.3.2	Failure time process model . . . . .	87
5.3.3	Baseline hazard estimation . . . . .	89
5.3.4	Probability of survival . . . . .	90
5.3.5	Evaluating prediction error . . . . .	91
5.4	Simulations . . . . .	93
5.5	Data Analysis: Cardiac care unit . . . . .	94
5.6	Discussion . . . . .	101
<b>6</b>	<b>Conclusion</b>	<b>103</b>
<b>A</b>	<b>Regularity conditions (Chapter 2)</b>	<b>109</b>
<b>B</b>	<b>Large sample theory (Chapter 2)</b>	<b>113</b>
B.1	Proof of Theorem 2.2.1 . . . . .	113
B.2	Proof that (2.20) is biased . . . . .	115
<b>C</b>	<b>GEE: extensions to other working correlation structures (Chapter 2)</b>	<b>117</b>
<b>D</b>	<b>Additional tables (Chapter 2)</b>	<b>119</b>
<b>E</b>	<b>Visitation assumptions (Chapter 3)</b>	<b>125</b>
<b>F</b>	<b>Theorems for consistency and asymptotic normality of solutions to es- timating equations (Chapters 3-5)</b>	<b>129</b>
F.1	Notation . . . . .	129
F.2	Consistency . . . . .	129
F.3	Asymptotic normality . . . . .	130
<b>G</b>	<b>Large sample theory (Chapter 3)</b>	<b>131</b>
G.1	Regularity conditions sufficient for Theorem 3.2.1 . . . . .	131
G.2	Proof of Theorem 3.2.1 . . . . .	132



G.3	Regularity conditions sufficient for Theorem 3.2.2 . . . . .	136
G.4	Proof of Theorem 3.2.2 . . . . .	137
<b>H</b>	<b>Missing data transition density (Chapter 4)</b>	<b>141</b>
<b>I</b>	<b>Conditional expectation calculations (Chapter 4)</b>	<b>143</b>
<b>J</b>	<b>Large sample theory (Chapter 4)</b>	<b>144</b>
J.1	Regularity conditions sufficient for Theorem 4.2.1 . . . . .	144
J.2	Proof of Theorem 4.2.1 . . . . .	145
<b>K</b>	<b>Large sample theory (Chapter 5)</b>	<b>149</b>
K.1	Regularity conditions sufficient for Theorem 5.3.1 . . . . .	149
K.2	Proof of Theorem 5.3.1 . . . . .	150

## LIST OF TABLES

2.1	Simulation Results for $n=50$ , $C_i \sim N(225, 0)$ . . . . .	21
2.2	Simulation Results for $n=50$ , $C_i \sim N(125, 0)$ . . . . .	22
2.3	Estimated regression coefficients and standard errors, Models 1 and 2 . . . . .	26
2.4	Estimated regression coefficients and standard errors, Model 3 . . . . .	28
3.1	Simulation results for (3.13) with mean reverting drift using a correctly specified (3.9) and one with a true $t_3$ diffusion variance . . . . .	49
3.2	Simulation results for (3.13) with deterministic drift using a correctly specified (3.9) and one with a true $t_3$ diffusion variance . . . . .	50
3.3	Marginal estimates of the intervention effects using the Lin-Ying method (no weighting), Buzkova-Lumley method, and the method proposed in this chapter with two types of weighting . . . . .	53
3.4	Conditional estimates of volatility of interventions . . . . .	53
3.5	Correlation parameter estimates based on (3.19) with their subscripted p-values for a two-sided test against zero . . . . .	58
3.6	Dynamical correlation parameter estimates for the Dubin and Müller (2005) model with their subscripted bootstrapped p-values for a two-sided test against zero . . . . .	58
4.1	Simulation results for mean reverting drift using (4.9). The three columns respectively represent no missingness, 30% missingness, and 30% missingness with a misspecified variance . . . . .	70
4.2	Simulation results for deterministic drift using (4.9). The three columns respectively represent no missingness, 30% missingness, and 30% missingness with a misspecified variance . . . . .	71
4.3	Estimate of $\kappa_0$ and its ASE in (4.13) for the CHF data . . . . .	77
4.4	Estimates of $\beta_0^{(1)}$ and $\beta_0^{(2)}$ , and their ASEs, from (4.14) for the CHF data. The second column corresponds to the use of assumption (3.8) with IPSW in (3.13); the final column corresponds to the use of assumption (4.1) without any IPSW in (3.13) . . . . .	78
5.1	Simulation results for deterministic drift using (5.4) with 400 subjects and both 50% censoring and 75% censoring . . . . .	94
5.2	Simulation results for mean reverting drift using (5.4) with 400 subjects and both 50% censoring and 75% censoring . . . . .	94
5.3	Time-dependent covariate definitions . . . . .	95
5.4	Time-fixed covariate definitions . . . . .	96
5.5	Prediction error estimates for different possible models . . . . .	98
D.1	Simulation results for $n=50$ , $C_i \sim N(225, 50)$ . . . . .	119
D.2	Simulation results for $n=50$ , $C_i \sim N(125, 50)$ . . . . .	120
D.3	Simulation results for $n=200$ , $C_i \sim N(225, 0)$ . . . . .	121

D.4	Simulation results for $n=200$ , $C_i \sim N(225, 50)$ . . . . .	122
D.5	Simulation results for $n=200$ , $C_i \sim N(125, 0)$ . . . . .	123
D.6	Simulation results for $n=200$ , $C_i \sim N(125, 50)$ . . . . .	124
E.1	A comparison of assumptions regarding the visitation process in five different papers and Chapter 3 of this thesis . . . . .	127

## LIST OF FIGURES

2.1	Distribution of numbers of recurrent events for drug and control subjects . . . . .	24
2.2	Standardized complete gap time residuals for Models 1-3 . . . .	29
3.1	Percentage homelessness across time for four randomly selected subjects. The conditional estimates at each visit time are connected with the solid blue line; the marginal estimates are connected with the dashed red line . . . . .	54
5.1	Baseline cumulative hazard of test times in the first 60 hours. The dotted vertical lines denote midnight. This model uses all the covariates from Tables 5.3 and 5.4 in the Cox model for visitation	97
5.2	ROC curves plotting sensitivity versus 1-specificity for different possible models . . . . .	99
5.3	Standardized residuals for covariates in the cardiac data . . . . .	100

# CHAPTER 1

## INTRODUCTION

Longitudinal studies often involve observations taken at irregular times. Sometimes interest lies in estimating the parameters governing the distribution of these observation times; other times, these parameters are considered a nuisance and the main goal is to model the process that is sampled at these discrete observation times, and possibly to also model a time-to-event that depends on the observation process and/or the longitudinal process. This thesis will consider the above situations using flexible methodologies which allow for censoring and the use of time-fixed and time-dependent covariates.

In Chapter 2 we directly model the conditional means and variances of the inter-event times, also known as gap times, of a particular recurrent event process using a generalized estimating equation (GEE) approach. Our method borrows from the literature on intensity models by allowing these gap times to depend on a complete history of observed information, and also borrows from the literature on marginal models by not fully specifying the within-subject covariance structure. The result is a robust method that takes the best attributes of both intensity and marginal models. Section 2.2 introduces the model and provides connections to the GEE literature and to Murphy et al. (1995). Simulation results are provided in Section 2.3, with extra tables in Appendix D, and an analysis of childhood asthma data, previously studied by Duchateau et al. (2003), is contained in Section 2.4. Section 2.5 concludes the chapter and outlines possible directions for future study. Technical details and large sample theory are provided in Appendices A-C. A version of this work, Clement and Strawderman (2009), was published in the journal *Biostatistics*, and an R pack-

age called `condGEE`, which encompasses our conditional GEE code, is available at the CRAN website.

Chapters 3, 4, and 5 consider the distribution of the observation times as a nuisance, and their goals are respectively to model a multivariate longitudinal process subject to censoring; to model a multivariate longitudinal process subject to censoring, failure, and missingness; and to model and predict a time-to-event based on the estimated longitudinal process.

These chapters were initially motivated by longitudinal observations on patients in a cardiac care unit, with the ultimate goal being prediction of survival beyond the current observation time. Much of the literature in this area considers jointly modeling the longitudinal and time-to-event processes (Wulfsohn and Tsiatis, 1997; Tsiatis and Davidian, 2001; Lin et al., 2002; Tsiatis and Davidian, 2004), but we consider a two step approach: Chapter 4 deals with estimation of parameters driving the longitudinal process, and thus can stand on its own, and Chapter 5 deals with estimation of parameters in the event time model. More specifically, these chapters proceed as follows.

Section 3.2 outlines a general methodology for the estimation of parameters indexing a multivariate longitudinal process observed only at discrete, informative visitation times, and subject to censoring. There are two particular cases of interest: modeling a response conditioned on covariates and modeling a response conditioned on covariates *and* its own past history; we call these approaches “marginal” and “conditional” respectively. The differences between the two approaches concern the types of questions they can answer, and the assumptions they make on the observation process. Section 3.3 provides simulation results for two possible conditional models, Section 3.4.1 re-analyzes

the HUD-VASH homelessness dataset, which was previously considered in Lin et al. (2004) and Bůžková and Lumley (2009), using our conditional and marginal estimating equations, and Section 3.4.2 studies longitudinal measurements of blood proteins relating to kidney function, a dataset which was first analyzed in Kaysen et al. (2000), and later in Dubin and Müller (2005). Section 3.5 wraps up the chapter and outlines possible future work. Details on observation process assumptions from a few different papers are provided in Appendix E and large sample theory is provided in Appendix G.

Section 4.2 introduces a model for intermittently missing responses, and also allows for an end of study due to a terminal event. This requires some new assumptions, foremost of which is that the response process follows a Gaussian distribution. Section 4.3 uses simulation to compare the resulting transition density parameter estimates to those from Chapter 3, where only censoring was considered. An analysis of medical cost data originally studied in Liu et al. (2008a) and Liu et al. (2008b) is provided in Section 4.4, and the chapter is wrapped up in Section 4.5. Some technical details are provided in Appendices H and I, and large sample theory is provided in Appendix J.

Chapter 5 changes the focus to the failure time process, and treats the longitudinal process as a covariate process used in failure prediction. This change from “response process” to “covariate process” is superficial, and the transition density parameters from Chapter 4 can be used to solve for the conditional distribution of covariates of an arbitrary subject at an arbitrary time - a calculation simplified by the previously assumed multivariate normality of the longitudinal process. This is done in Section 5.2. Section 5.3 then proposes two models based on the Cox model hazard that make use of the full estimated covariate process

trajectories to predict survival beyond the current observation time. Proposed models can then be compared based on their prediction errors (Schoop et al., 2008; Schoop, 2008). More simulations are carried out in Section 5.4 and an analysis of cardiac data is provided in Section 5.5. These data were provided to us by Dr. Mark Cowen in Ann Arbor, Michigan. Finally, Section 5.6 concludes by discussing the contributions presented in the previous sections, and proposing future directions of study. Large sample theory is provided in Appendix K.

Chapter 6 reviews all the contributions of this thesis and in particular highlights possible directions of study that build on the work presented here.



## CHAPTER 2

### CONDITIONAL GEE FOR RECURRENT EVENT GAP TIMES

#### 2.1 Introduction

In longitudinal studies, each subject may experience several consecutive events of the same basic type. Such “recurrent event” outcome data now constitute a heavily studied area in statistics, with applications ranging from economics to engineering to biomedicine. Examples common in medicine and public health applications include recurrent infections and other diseases, hospitalizations, and seizures. The development of useful regression models for recurrent outcome data is therefore a problem of significant practical and methodological interest. Beginning with the extension of the proportional hazards regression model of Cox (1972) to the case of multivariate counting processes by Andersen and Gill (1982), a significant literature on this topic has developed over the past 25 years. In general, methods for analyzing recurrent event data can be cross-classified into one of four categories determined by: (i) the choice of “calendar” versus “gap” times as the fundamental temporal scale; and, (ii) the use of “marginal” versus “intensity” models for analyzing the data. Several classes of marginal and intensity models have been proposed for analyzing recurrent event outcomes on each time scale; an extensive, contemporary review of existing methods is available in Cook and Lawless (2007).

When the events are all considered to be of the same type, the gap time scale is arguably the most natural and informative time scale for analysis. Examples of marginal models developed for the gap time setting include Chang and Wang (1999); Chang (2004); Chen and Wang (2004); Huang (2002); Huang and Chen

(2003); Lin et al. (1999); Prentice et al. (1981). Gap-time focused intensity models are considered in Aalen and Husebye (1991); Oakes and Cui (1994); Peña et al. (2001); Duchateau et al. (2003); Strawderman (2005) and Strawderman (2006), among others. In an interesting paper, Murphy, Bentley, and O’Hanesian (1995) propose a methodology that is difficult to wholly classify into a single category. Specifically, Murphy et al. (1995) introduce a variant of the estimating equations considered in Murphy and Li (1995) in order to develop a model appropriate for describing the conditional mean length of women’s menstrual cycles. In this model, the mean of the current cycle is allowed to depend on aspects of past cycle behavior and/or other time-fixed and time-dependent covariates. The methodology developed in Murphy et al. (1995) forms a starting point for developing a large and useful class of methods for conducting gap time analyses that relaxes the stringent restrictions imposed by simpler marginal models while avoiding the need to fully specify how the probability of subsequent recurrence depends on the prior event and covariate histories.

This chapter expands upon the work of Murphy et al. (1995) in several useful ways, including: making allowances for transformations of gap times, providing clarification of the censoring conditions under which the proposed estimating equations are unbiased, correcting two important errors in the original paper, further illuminating the connections to generalized estimating equations (GEE), and the provision of an appropriate large sample theory. The chapter will proceed as follows. The proposed methodology is first developed in Sections 2.2.1-2.2.3; Sections 2.2.4 and 2.2.5 then respectively establish the connections to the methodology originally proposed in Murphy et al. (1995) and to GEE (Liang and Zeger, 1986). Section 2.3 contains an expanded version of the simulation study summarized in Tables II and III of Murphy et al. (1995); extra tables are

found in Appendix D. In Section 2.4, the proposed methodology is used to re-analyze the asthma prevention trial data of Duchateau et al. (2003). We close the chapter in Section 2.5 with a discussion of several interesting directions for further study. Technical conditions and large sample theory are provided in Appendices A-C.

## 2.2 Methodology

### 2.2.1 Notation and model

For simplicity, we introduce the notation for just one subject; the inclusion of an additional subscript permits immediate extension to the case of multiple subjects, as will be required in Section 2.2.3. We assume that the time origin for analysis is  $S_0 = 0$ , with subsequent events occurring at times  $0 < S_1 < S_2 < \dots$  until observation terminates at an observed time  $C > 0$ . It is assumed that  $S_1$  represents a complete observation time, thereby covering those cases in which observation begins with the occurrence of an event. In settings where observation starts in between two events,  $S_0$  is taken to represent the time of the first event subsequent to the start of observation. Despite a mild loss of information, such a convention does not cause bias provided that the decision to delete this initial time period is made independently of its length (e.g. Aalen and Husebye, 1991; Murphy et al., 1995).

Let  $N(u) = \max\{n \geq 1 : S_n \leq u\}$  count the number of events up to and including time  $u$ , and let  $N = N(C)$ . The observed data on this subject is assumed to

take the form

$$\mathbb{O} = \{S_1, S_2, \dots, S_N; \bar{L}_1, \dots, \bar{L}_{N+1}; C\}, \quad (2.1)$$

where  $\bar{L}_j$  denotes the covariate information available at time  $S_{j-1}$  for  $j \geq 1$  and may include baseline covariates, covariates measured at or before each event time, and summaries of the past event and covariate history. It is not assumed that  $\bar{L}_j$  necessarily captures the full covariate or event history up to time  $S_{j-1}$ . Let  $X_j = S_j - S_{j-1}$  denote the  $j^{\text{th}}$  gap time and define  $Y_j = h(X_j)$ , where  $h(\cdot)$  is a specified monotone nondecreasing transformation. Also, let  $\mathcal{H}_j = \{S_1, \dots, S_{j-1}; \bar{L}_1, \dots, \bar{L}_j\}$  denote the cumulative information concerning the event and covariate histories assumed available through time  $S_{j-1}$ ; note that  $\mathcal{H}_j \subset \mathcal{H}_{j+1}$  for  $j \geq 1$ .

The fundamental modeling assumption of this chapter is

$$E[Y_j | \mathcal{H}_j] = \mu_j(\theta) \quad \text{and} \quad \text{Var}[Y_j | \mathcal{H}_j] = \sigma^2 V_j^2(\theta), \quad j \geq 1, \quad (2.2)$$

where  $\mu_j(\theta) \in \mathbb{R}$  and  $V_j(\theta) > 0$  are known scalar functions of the parameter vector  $\theta$  and  $\sigma^2 > 0$ . Importantly, these means and variances are defined conditionally upon  $\mathcal{H}_j$  and are therefore independent of censoring information. Further assumptions are needed in order to properly deal with the presence of censoring; see, for example, Sections 2.2.2 and 2.2.3 as well as Conditions (A0) and (A1) of Appendix A. Three examples of interesting model choices include

$$Y_j = X_j \text{ for } j \geq 1, E[Y_j | \mathcal{H}_j] = \mu_j(\theta) \quad \text{and} \quad \text{Var}[Y_j | \mathcal{H}_j] = \sigma^2 V_j^2(\theta); \quad (2.3)$$

$$Y_j = X_j \text{ for } j \geq 1, E[Y_j | \mathcal{H}_j] = \mu_j(\theta) \quad \text{and} \quad \text{Var}[Y_j | \mathcal{H}_j] = \sigma^2 \mu_j^2(\theta); \quad (2.4)$$

$$Y_j = \log(X_j) \text{ for } j \geq 1, E[Y_j | \mathcal{H}_j] = \mu_j(\theta) \quad \text{and} \quad \text{Var}[Y_j | \mathcal{H}_j] = \sigma^2 V_j^2(\theta). \quad (2.5)$$

Murphy et al. (1995) consider the model (2.3), allowing for the possibility of a general variance function  $V_j^2(\theta)$ . Models (2.4) and (2.5) provide useful general-

izations of the Accelerated Gap Times (AGT) model proposed in Strawderman (2005). The AGT model assumes that the gap times of the recurrent event process satisfy  $X_j = R_j\mu(\theta)$ , where  $\{R_j, j \geq 1\}$  are independent and identically distributed with a distribution independent of  $\theta$  and  $\mu(\theta)$  accelerates or decelerates the baseline gap times for a subject based on the values of time-independent covariates. Assuming that  $E[R_j] = 1$ , model (2.4) is observed to be a direct generalization of this model upon taking  $\mu_j(\theta) = \mu(\theta)$ ,  $V_j(\theta) = \mu(\theta)$ , and  $\sigma^2 = \text{Var}[R_j]$  for  $j \geq 1$ . Taking logs, the alternative model  $\log X_j = \log \mu(\theta) + \log R_j$  is obtained; (2.5) evidently covers this form of the AGT model with  $\mu_j(\theta) = \log \mu(\theta)$ ,  $E[\log R_j] = 0$ ,  $\sigma^2 = \text{Var}[\log R_j]$ , and  $V_j(\theta) = 1$  for  $j \geq 1$ .

## 2.2.2 Estimation with “full” and observed data: single subject

For convenience, define for  $j \geq 1$  the following notation:

$$f_j(\theta) = \frac{d\mu_j(\theta)}{d\theta} V_j^{-1}(\theta) \quad \text{and} \quad Z_j(\theta) = \frac{Y_j - \mu_j(\theta)}{V_j(\theta)}. \quad (2.6)$$

For simplicity, we make no distinction between  $\eta = (\theta^T, \sigma^2)^T$  and the data generating parameter  $\eta_0$  throughout this section. Under (2.2), it follows that  $E[Z_j(\theta)|\mathcal{H}_j] = 0$  and  $E[Z_j^2(\theta)|\mathcal{H}_j] = \sigma^2$  for each  $j \geq 1$ . A naïve approach to the estimation of  $\eta$  might therefore begin with consideration of the estimating equations

$$\sum_{j=1}^N f_j(\theta) Z_j(\theta) \quad \text{and} \quad \sum_{j=1}^N b_j(\eta) (Z_j^2(\theta) - \sigma^2),$$

where  $b_j(\eta)$  is a scalar weight function satisfying  $E[b_j(\eta)|\mathcal{H}_j] = b_j(\eta)$  for each  $j \geq 1$ . However, these estimating equations generally fail to be unbiased because each utilizes only the complete gap times  $X_1, \dots, X_N$ , each of which satisfies  $X_j \leq C, j = 1, \dots, N$ .

Let  $\mathbb{F} = \{S_1, S_2, \dots, S_N, S_{N+1}; \bar{L}_1, \dots, \bar{L}_{N+1}; C\}$  denote the observed data (2.1), augmented with the additional information on the first event time  $S_{N+1}$  following time  $C$ . Since  $S_{N+1}$  is not generally observable, one may view  $\mathbb{F}$  as a suitable representation of “full data” in this setting. Consider the pair of  $\mathbb{F}$ –dependent estimating equations (cf. Murphy and Li, 1995)

$$D_{F,1}^*(\eta) \equiv \sum_{j=1}^{N+1} f_j(\theta) Z_j(\theta) \quad \text{and} \quad D_{F,2}^*(\eta) \equiv \sum_{j=1}^{N+1} b_j(\eta) (Z_j^2(\theta) - \sigma^2). \quad (2.7)$$

Theorem 2.2.1, proved in Appendix B.1, shows that the  $\mathbb{F}$ –dependent estimating equations (2.7) are unbiased under Condition (A0) of Appendix A.

**Theorem 2.2.1.** Under Condition (A0), the estimating equations (2.7) are unbiased.

While unbiased, the estimating equations (2.7) cannot be used directly for estimating  $\eta$  from the observed data (2.1) because each depends on  $Y_{N+1} = h(X_{N+1})$ , information not available under (2.1). Similarly to Murphy et al. (1995), one starting point for developing practically useful estimating equations is to project (2.7) onto the observed data:

$$E[D_{F,1}^*(\eta)|\mathbb{O}] = \sum_{j=1}^N f_j(\theta) Z_j(\theta) + f_{N+1}(\theta) E[Z_{N+1}(\theta)|\mathbb{O}] \quad (2.8)$$

and

$$E[D_{F,2}^*(\eta)|\mathbb{O}] = \sum_{j=1}^N b_j(\eta) (Z_j^2(\theta) - \sigma^2) + b_{N+1}(\eta) E[(Z_{N+1}^2(\theta) - \sigma^2)|\mathbb{O}]. \quad (2.9)$$

Using iterated expectation, an easy calculation shows that the  $\mathbb{O}$ –dependent estimating equations (2.8) and (2.9) remain unbiased under the conditions of Theorem 2.2.1.

Defining  $W_j(\eta) = \sigma^{-1}Z_j(\theta)$  and  $H_j = \mathcal{H}_j \cup \{C \geq S_{j-1}\}$  for  $j \geq 1$ , the expectations appearing on the right-hand side of (2.8) and (2.9) can be rewritten as follows:

$$\begin{aligned} E[Z_{N+1}(\theta)|\mathbb{O}] &= \sigma E[W_{N+1}(\eta)|W_{N+1}(\eta) > w(\eta), H_{N+1}], \\ E[Z_{N+1}^2(\theta) - \sigma^2|\mathbb{O}] &= \sigma^2 \left( E[W_{N+1}^2(\eta)|W_{N+1}(\eta) > w(\eta), H_{N+1}] - 1 \right), \end{aligned}$$

where  $w(\eta) = [\sigma V_{N+1}(\theta)]^{-1} \{h(C - S_N) - \mu_{N+1}(\theta)\}$  is considered fixed in each expression.

The dependence of  $W_{N+1}(\eta)$  on  $Y_{N+1}$  implies that  $W_{N+1}(\eta)$  represents missing data under (2.1). The projections (2.8) and (2.9) therefore correspond to using an obvious form of conditional imputation and further modeling assumptions that permit computation of these conditional expectations are needed. Under Condition (A0),  $Y_{N+1}$  and hence  $W_{N+1}(\eta)$  may be considered missing at random (MAR, Rubin, 1976) and immediate progress is possible under a parametric specification for the conditional distribution  $W_{N+1}(\eta)|H_{N+1}$ . Noting that  $\{W_j(\eta), j \geq 1\}$  is a sequence of dependent, standardized (i.e., mean zero, variance one) random variables, we propose to proceed under the simplifying assumption that  $W_{N+1}(\eta)|H_{N+1}$  is distributed according to a fully specified parametric distribution  $F_0(\cdot)$  having mean zero and variance one. This immediately implies that

$$E[W_{N+1}^r(\eta)|W_{N+1}(\eta) > w(\eta), H_{N+1}] = K_r(w(\eta)), \quad r = 1, 2, \quad (2.10)$$

where

$$K_r(w) = \int_w^\infty u^r \frac{dF_0(u)}{1 - F_0(w-)}. \quad (2.11)$$

Under (2.10) and (2.11), (2.8) and (2.9) reduce to

$$E[D_{F,1}^*(\eta)|\mathbb{O}] = \sum_{j=1}^N f_j(\theta)Z_j(\theta) + \sigma f_{N+1}(\theta)K_1(w(\eta)) \quad (2.12)$$

and

$$E[D_{F,2}^*(\eta)|\mathbb{O}] = \sum_{j=1}^N b_j(\eta) \left( Z_j^2(\theta) - \sigma^2 \right) + \sigma^2 b_{N+1}(\eta) (K_2(w(\eta)) - 1). \quad (2.13)$$

For example, with  $F_0(x) = \Phi(x)$ ,

$$K_1(x) = \frac{\phi(x)}{1 - \Phi(x)} \quad \text{and} \quad K_2(x) = 1 + \frac{x\phi(x)}{1 - \Phi(x)},$$

whereas the choice  $F_0(x) = 1 - e^{-(x+1)}$  for  $x > -1$  leads to  $K_1(x) = x + 1$  and  $K_2(x) = 1 + (x + 1)^2$ . Each of (2.12) and (2.13) has mean zero provided (2.10) holds and  $F_0(\cdot)$  is correctly specified.

The moment assumptions on  $F_0(\cdot)$  are natural in view of the mean-variance specification of the model. In addition, this parsimonious model facilitates straightforward implementation and justification of inference procedures, as will be seen in Section 2.2.3. We emphasize here that the parametric specification of  $F_0(\cdot)$  is only introduced for the purposes of dealing with the censored time  $Y_{N+1} = h(X_{N+1})$ . We have not assumed that each member of the sequence  $\{W_j(\eta), j \geq 1\}$  has distribution  $F_0(\cdot)$ ; in addition, we have not introduced any assumptions that impose a fully parametric dependence structure on  $\{W_j(\eta), j \geq 1\}$ . Perhaps the easiest way to see this is to note that such distributional assumptions are neither needed nor used in the development of (2.7); see also Murphy and Li (1995) for related results and discussion.

REMARK: One may write  $K_1(x) = E[W|W > x] = x + M(x)$ , where  $W$  is a random variable with distribution function  $F_0(\cdot)$  and  $M(x) = E[W - x|W > x]$  is the corresponding mean residual life function. An immediate consequence of this relationship is that one cannot specify a valid parametric model for  $K_1(\cdot)$  without also specifying a valid parametric model for  $F_0(\cdot)$  (e.g., Cox, 1962; Oakes and Dasu, 1990; Kotz and Shanbhag, 1980).



### 2.2.3 Estimation and inference with observed data: $n$ subjects

Suppose there is data available on  $n$  independent subjects, say  $\mathbb{O}_i$ ,  $i = 1, \dots, n$ , where  $\mathbb{O}_i$  is the data (2.1) on the  $i^{\text{th}}$  subject. Then, (2.12) and (2.13) immediately generalize, yielding the pair of estimating equations

$$S_{n,1}(\eta) = \frac{1}{n} \sum_{i=1}^n \left\{ \sum_{j=1}^{N_i} f_{ij}(\theta) W_{ij}(\eta) + f_{i,N_i+1}(\theta) K_1(w_i(\eta)) \right\} \quad (2.14)$$

and

$$S_{n,2}(\eta) = \frac{1}{n} \sum_{i=1}^n \left\{ \sum_{j=1}^{N_i} b_{ij}(\eta) (W_{ij}^2(\eta) - 1) + b_{i,N_i+1}(\eta) (K_2(w_i(\eta)) - 1) \right\}, \quad (2.15)$$

where

$$W_{ij}(\eta) = \frac{Y_{ij} - \mu_{ij}(\theta)}{\sigma V_{ij}(\theta)} \quad \text{and} \quad w_i(\eta) = \frac{h(C_i - S_{iN_i}) - \mu_{i,N_i+1}(\theta)}{\sigma V_{i,N_i+1}(\theta)}.$$

Assuming that (2.10) holds and  $F_0(\cdot)$  has been correctly specified, (2.14) and (2.15) together form a collection of unbiased estimating equations for  $\eta$ . In order to solve these equations and expect to obtain a unique solution  $\widehat{\eta}_n = (\widehat{\theta}_n^T, \widehat{\sigma}_n^2)^T$  for finite  $n$ , smoothness assumptions on  $K_r(\cdot)$ ,  $r = 1, 2$  are required. Conditions (A3)-(A5) of Appendix A impose sufficient smoothness restrictions on  $K_r(\cdot)$ ,  $r = 1, 2$ ; see Appendix A for further details and discussion.

Let  $S_n(\eta) = (S_{n,1}(\eta)^T, S_{n,2}(\eta)^T)^T$ ; then, we may write

$$S_n(\eta) = \frac{1}{n} \sum_{i=1}^n \psi(\eta, \mathbb{O}_i), \quad (2.16)$$

where  $\psi(\eta, \mathbb{O}_i) = (\psi_1(\eta, \mathbb{O}_i)^T, \psi_2(\eta, \mathbb{O}_i)^T)^T$  is a vector of known functions of  $\eta$  and  $\mathbb{O}_i$ ,  $i = 1, \dots, n$ . Define  $S(\eta) = E_{\eta_0}[\psi(\eta, \mathbb{O}_1)]$  and  $S'(\eta) = \frac{d}{d\eta} S(\eta)$ . Theorems 2.2.2 and 2.2.3 show that  $\widehat{\eta}_n$  is both consistent and asymptotically normal as  $n \rightarrow \infty$ ; see Appendix A for the statement of regularity conditions and further details on proof.

**Theorem 2.2.2.** Under conditions (A0)-(A4) and as  $n \rightarrow \infty$ , there exists a sequence  $\{\widehat{\eta}_n, n \geq 1\}$  and a unique  $\eta_0$  such that  $S(\eta_0) = 0$ ,  $S_n(\widehat{\eta}_n) = 0$  with probability going to one, and  $\widehat{\eta}_n \xrightarrow{P} \eta_0$ .

**Theorem 2.2.3.** Under conditions (A0)-(A7), the sequence  $\sqrt{n}(\widehat{\eta}_n - \eta_0)$  is asymptotically normal with mean zero and covariance matrix

$$S'(\eta_0)^{-1} E_{\eta_0} [\psi(\eta_0, \mathbb{O}_1) \psi(\eta_0, \mathbb{O}_1)^T] (S'(\eta_0)^{-1})^T. \quad (2.17)$$

It further follows that one can consistently estimate  $\text{Var}(\widehat{\eta}_n)$  via

$$n^{-1} S'_n(\widehat{\eta}_n)^{-1} \left( \frac{1}{n} \sum_{i=1}^n \psi(\widehat{\eta}_n, \mathbb{O}_i) \psi(\widehat{\eta}_n, \mathbb{O}_i)^T \right) (S'_n(\widehat{\eta}_n)^{-1})^T. \quad (2.18)$$

REMARK: The asymptotic results for  $\widehat{\eta}_n$  rely on the assumption that (2.10) holds with  $F_0(\cdot)$  correctly specified. As pointed out earlier in Section 2.2.2, the parametric imputation assumption (2.10) has been introduced in order to deal with the censoring of  $Y_{i,N_i+1} = h(X_{i,N_i+1})$ ,  $i = 1, \dots, n$ . Other models for imputation, as well as methods for handling the missing data problem, are possible. However, under a MAR specification, all such methods rely on further modeling assumptions and, similarly to the problem of misspecifying  $F_0(\cdot)$ , incorrect specifications create bias. We refer the reader to Section 2.5 for further discussion.

## 2.2.4 Connections to Murphy et al. (1995)

Similarly to Section 2.2.3, Murphy et al. (1995) focus on estimating  $\eta$  from the observed data  $\mathbb{O}_i$ ,  $i = 1, \dots, n$  when  $h(x) = x$ , suggesting an adaptation of the EM

algorithm of Dempster et al. (1977) for use with estimating equations, an idea recently explored in greater generality by Elashoff and Ryan (2004). Specifically, for each  $i = 1, \dots, n$  and given current estimates  $\widehat{\theta}^{(k)}$  and  $\widehat{\sigma}^{(k)}$  of  $\theta$  and  $\sigma$ , Murphy et al. (1995) suggest imputing

$$Y_{i,N_i+1,r}^{(k)} = \mu_{i,N_i+1}(\widehat{\theta}^{(k)}) + \widehat{\sigma}^{(k)} V_{i,N_i+1}(\widehat{\theta}^{(k)}) \epsilon_r, \quad (2.19)$$

where  $\epsilon_r, r = 1, \dots, B$  are independent and identically distributed mean zero, variance one random variables. These  $B$  variables are then used to compute both  $\widetilde{E}_i^{(k)} = \sum_{r=1}^B Y_{i,N_i+1,r}^{(k)} w_{ir}^{(k)}$  and  $\widetilde{V}_i^{(k)} = \sum_{r=1}^B (Y_{i,N_i+1,r}^{(k)} - \widetilde{E}_i^{(k)})^2 w_{ir}^{(k)}$ , where

$$w_{ir}^{(k)} = \frac{I(Y_{i,N_i+1,r}^{(k)} > C_i - S_{iN_i})}{\sum_{s=1}^B I(Y_{i,N_i+1,s}^{(k)} > C_i - S_{iN_i})}, \quad i = 1, \dots, n.$$

The estimates  $\widehat{\theta}^{(k)}$  and  $\widehat{\sigma}^{(k)}$  are then updated according to the following procedure. Expressed in our notation,  $\widehat{\theta}^{(k+1)}$  is first computed by solving

$$S_M(\theta) = \frac{1}{n} \sum_{i=1}^n \left\{ \sum_{j=1}^{N_i} f_{ij}(\theta) Z_{ij}(\theta) + f_{i,N_i+1}(\theta) \left[ \frac{\widetilde{E}_i^{(k)} - \mu_{i,N_i+1}(\theta)}{V_{i,N_i+1}(\theta)} \right] \right\} = 0.$$

As indicated in the Appendix of Murphy et al. (1995), one then computes

$$\widehat{\sigma}^{(k+1)} = \left( \frac{1}{n + \sum_{i=1}^n N_i} \sum_{i=1}^n \left\{ \sum_{j=1}^{N_i} Z_{ij}^2(\widehat{\theta}^{(k+1)}) + \frac{\widetilde{V}_i^{(k)}}{[V_{i,N_i+1}(\widehat{\theta}^{(k+1)})]^2} \right\} \right)^{1/2}. \quad (2.20)$$

This iteration continues until the relative change in each estimated model parameter is small.

We now illuminate the connections to the methodology summarized in Section 2.2.3, as well as an important problem with the methodology described above. The “E-type” step described above involves computing Monte Carlo approximations to the conditional cumulants  $E[Y_{i,N_i+1} | Y_{i,N_i+1} > C_i - S_{iN_i}, H_{i,N_i+1}]$

and  $\text{Var}[Y_{i,N_i+1}|Y_{i,N_i+1} > C_i - S_{iN_i}, H_{i,N_i+1}]$ , evaluated at the current parameter values  $\widehat{\theta}^{(k)}$  and  $\widehat{\sigma}^{(k)}$ . Under the imputation assumption (2.19), it is apparent that  $\widetilde{E}_i^{(k)}$  is a Monte Carlo approximation to  $\sigma K_1(C_i - S_{iN_i})$   $i = 1, \dots, n$ , where  $K_1(\cdot)$  is defined in (2.11) and  $\epsilon_r \sim F_0$ ,  $r = 1, \dots, B$ . The use of (2.14) in place of the Monte Carlo approximation  $S_M(\theta)$  used by Murphy et al. (1995) reduces computational demands and leads to a stable estimation procedure independent of Monte Carlo error. The use of (2.15) corrects a fundamental error in Murphy et al. (1995). Specifically, as shown in Appendix B.2, the estimator (2.20) is biased. Moreover, the degree of bias increases with fewer complete observations per subject because the censored cycles contribute an increased proportion of the information to the estimating equation.

REMARK: A less subtle error corrected by this thesis involves the variance estimate (2.18). The corresponding estimate proposed in Murphy et al. (1995, Appendix) assumes independence of the gap times within subjects and is therefore inconsistent when this assumption fails.

## 2.2.5 Connections with GEE

Similarly to Section 2.2.2, define  $\mathbb{F}_i = \{S_{i1}, S_{i2}, \dots, S_{iN_i}, S_{i,N_i+1}; \bar{L}_{i1}, \dots, \bar{L}_{i,N_i+1}; C_i\}$ ,  $i = 1, \dots, n$ . Then, assuming  $\mathbb{F}_i$ ,  $i = 1, \dots, n$  represents the available data and also that subjects are independent of each other, the results of Theorem 2.2.1 imply that

$$U_{n,1}(\eta) = \frac{1}{n} \sum_{i=1}^n \left\{ \sum_{j=1}^{N_i+1} f_{ij}(\theta) W_{ij}(\eta) \right\} \quad (2.21)$$

and

$$U_{n,2}(\eta) = \frac{1}{n} \sum_{i=1}^n \left\{ \sum_{j=1}^{N_i+1} b_{ij}(\eta) (W_{ij}^2(\eta) - 1) \right\} \quad (2.22)$$

form a system of unbiased (full data) estimating equations for  $\eta$ . Under the same assumptions leading to (2.14) and (2.15), it follows that the projections of (2.21) and (2.22) onto the observed data  $\mathbb{O}_i$ ,  $i = 1, \dots, n$  reproduce (2.14) and (2.15).

The estimating equations (2.21) and (2.22) represent a particular example of a “full data” GEE system. Specifically, define for  $i = 1, \dots, n$  the matrices

$$A_i(\theta) = \text{diag}\{V_{i1}(\theta), \dots, V_{i,N_i+1}(\theta)\} \quad \text{and} \quad G_i(\theta) = \begin{pmatrix} \frac{d\mu_{i1}(\theta)}{d\theta} & \frac{d\mu_{i2}(\theta)}{d\theta} & \dots & \frac{d\mu_{i,N_i+1}(\theta)}{d\theta} \end{pmatrix}.$$

In addition, let  $I_{N_i+1}$  denote the identity matrix of dimension of  $(N_i + 1) \times (N_i + 1)$ ,  $i = 1, \dots, n$ . Then, one may write  $U_{n,1}(\eta)$  as

$$(n\sigma)^{-1} \sum_{i=1}^n G_i(\theta) [A_i(\theta) I_{N_i+1} A_i(\theta)]^{-1} \epsilon_i(\theta),$$

where  $\epsilon_i(\theta)$  is a vector with elements  $Y_{ik} - \mu_{ik}(\theta)$ ,  $k = 1, \dots, N_i + 1$ . The correspondence between (2.21) and a GEE system is now evident, the use of  $I_{N_i+1}$  in  $[A_i(\theta) I_{N_i+1} A_i(\theta)]^{-1}$  further imposing a “working independence” correlation structure on  $Y_{i1}, \dots, Y_{i,N_i+1}$  (cf. Molenberghs and Verbeke, 2005, Sec. 8.2). A similar construction is possible for (2.22); moreover, (2.21) and (2.22) together form a particular example of a GEE2 (e.g. Prentice and Zhao, 1991) system that imposes a block diagonal covariance structure on  $U_n(\eta) = (U_{n,1}(\eta)^T, U_{n,2}(\eta))^T$  (cf. Molenberghs and Verbeke, 2005, Sec. 8.5). In Appendix C, we consider the use of alternative working correlation structures, demonstrating in particular that valid structures must respect the conditional specification of the model in order for (2.21) and (2.22) to remain unbiased.

## 2.3 Simulations

The work of Murphy et al. (1995) was motivated by the analysis of menstrual cycle patterns. More specifically, the authors were interested in developing insight into the relationship between cycle length and covariates such as location, body mass index (BMI) and age. Murphy et al. (1995) also carried out a small simulation study modeled after these data in order to evaluate the robustness of their methods to assumptions regarding the nature of the censored cycle length. The following conditional mean and variance specifications were respectively used in both the analysis and simulation study:  $E[Y_{ij}|\mathcal{H}_{ij}] = \mu_{ij}(\theta)$ , where

$$\mu_{ij}(\theta) = \bar{L}_{ij}^T \gamma + \frac{\rho}{\rho(j-1) + 1 - \rho} \left[ \sum_{l=1}^{j-1} Y_{il} - \sum_{l=1}^{j-1} \bar{L}_{il}^T \gamma \right] \quad (2.23)$$

and  $\theta = (\gamma^T, \rho)^T$ ; and,  $\text{Var}[Y_{ij}|\mathcal{H}_{ij}] = \sigma^2 V_j^2(\theta)$ , where

$$V_{ij}(\theta) = \left( \left| 1 + \frac{\rho}{\rho(j-1) + 1 - \rho} \right| \right)^{1/2}. \quad (2.24)$$

The use of (2.23) and (2.24) evidently corresponds to a special case of (2.3). Assuming an equal number of cycles per woman, the proposed mean function corresponds to a simple linear mixed effects model using woman as a random effect. The parameter  $\rho$  can thus be interpreted as an intra-woman correlation coefficient, with  $\rho = 0$  denoting that past cycle lengths are not useful in predicting current cycle length.

In this section, we consider an expanded version of the simulations considered in Murphy et al. (1995). In addition to allowing for the possibility of a misspecified censored cycle length distribution, we consider the impact of both varying and shorter observation periods (i.e., fewer observed events), varying number of subjects, and two specifications of  $V_{ij}(\theta)$ . More specifically, cycle

data on either 50 or 200 independent women are simulated by generating cycle times according to the model  $Y_{ij} = \max\{Y_{ij}^*, 1\}$ ,  $j \geq 1$ ,  $i = 1, \dots, 50$ , where  $Y_{ij}^* = \mu_{ij}(\theta) + \sigma V_{ij}(\theta) \epsilon_{ij}$  and  $\epsilon_{ij}$  are independent, identically distributed observations from either a standard normal density or a shifted exponential density with mean zero and unit variance. As in Murphy et al. (1995), we assume that  $\mu_{ij}(\theta)$  is specified according to the following trivial modification of (2.23):

$$\mu_{ij}(\theta) = 28 + \gamma_0 + \gamma_1 \overline{\text{BMI}}_{ij} + \frac{\rho}{\rho(j-1) + 1 - \rho} \left[ \sum_{l=1}^{j-1} Y_{il} - \sum_{l=1}^{j-1} \{28 + \gamma_0 + \gamma_1 \overline{\text{BMI}}_{il}\} \right], \quad (2.25)$$

where  $\gamma_0 = 0.6$ ,  $\gamma_1 = -0.4$ , and  $\rho = 0.03$ . The single time-dependent covariate  $\overline{\text{BMI}}_{ij} = \text{BMI}_{ij} - 21$ , where  $\text{BMI}_{ij}$  is assumed to decrease linearly from  $22 \text{ kg/m}^2$  on day 1 to  $20 \text{ kg/m}^2$  on day 195, increase linearly back to  $21 \text{ kg/m}^2$  on day 225, and then remain constant thereafter. As in Murphy et al. (1995), we consider the specification (2.24) for  $V_{ij}(\theta)$  in conjunction with  $\sigma^2 = 11$ ; in addition, we also consider the specification  $V_{ij}(\theta) = |\mu_{ij}(\theta)|$ , where  $\mu_{ij}(\theta)$  is given in (2.25) and  $\sigma^2 = 1/72.2 = 0.014$ . These two choices of  $V_{ij}(\theta)$  correspond to the model specifications (2.3) and (2.4); our choices of  $\sigma^2$  approximately equalize the variances of the average gap times across the two settings. Finally, we assume the observation period for subject  $i$  is  $[0, C_i]$ , where  $C_i = \max\{C_i^*, 1\}$  and  $C_i^*$  is normally distributed. Four possible settings are considered, with  $E[C_i^*]$  set to either 125 or 225 days and  $\text{Var}(C_i^*)$  set either 0 or 50, respectively. The average number of events per subject under an observation period with  $E[C_i^*] = 125$  ( $E[C_i^*] = 225$ ) is approximately 3.9 (7.4). All simulations were run using  $b_{ij}(\eta) = 1$ , a choice that corresponds to the use of generalized least squares.

Tables 2.1 and 2.2 respectively summarize the results for  $C_i^* \sim N(225, 0)$  and  $C_i^* \sim N(125, 0)$ ; a comparison of these tables demonstrates the impact of the expected number of events. The remaining simulation results are summarized in

Tables D.1-D.6 in Appendix D. The top panel of Table 2.1 corresponds to the simulations summarized in Table II of Murphy et al. (1995). Each table corresponds to one combination of censoring distribution and sample size and summarizes the results for the four possible combinations of true and assumed error distributions for each choice of variance function (i.e.,  $V_{ij}(\theta)$ ). In the tables,  $|\text{rBias}|$  is the absolute relative bias of the average estimate and ESE is the empirical standard deviation of the estimate; both are computed from 1000 simulated datasets of the indicated sample size. The average estimated asymptotic standard errors (ASE) are obtained by averaging the square root of the diagonal of (2.18) over the 1000 simulated datasets. Relative bias is reported because the magnitude of  $\sigma^2$  differs greatly across the top and bottom halves of each table.

In general, the simulation results demonstrate that the model specifications (2.3) and (2.4) produce estimates of  $\theta$  and  $\sigma^2$  with comparable relative biases. The standard errors for  $\widehat{\theta}_n$  are also comparable across models and ASE provides an acceptable approximation to ESE in all cases. While a strong effect of misspecifying  $F_0(\cdot)$  is absent, the impact of doing so on  $\rho$  and  $\sigma^2$  does become more apparent when the sample size is increased to  $n = 200$ . In addition, comparing the results for Tables D.3 & D.5 (Appendix D), one can additionally see that biases increase under both incorrect and correct model misspecifications when the average number of expected events decreases. In general, though, the relative biases remain modest in most cases and the signs of all estimated parameters are also correct (results not shown).

Other interesting patterns also arise in these tables. For example, the bias of  $\rho$  is typically the largest, especially so when the true  $F_0(\cdot)$  is normally distributed. In addition, the sign of the bias is frequently negative (results not



Table 2.1: Simulation Results for  $n=50$ ,  $C_i \sim N(225, 0)$ 

Expected number of events $\doteq 7.4$			True $F_0$					
Model	Imputed $F_0$	Parameter	Normal			Exponential		
			rBias	ESE	ASE	rBias	ESE	ASE
$\mu_{ij}(\theta) = (2.25)$ $V_{ij}(\theta) = (2.24)$	Normal	$\gamma_0$	0.003	0.193	0.188	0.009	0.189	0.191
		$\gamma_1$	0.014	0.276	0.272	0.034	0.267	0.260
		$\rho$	0.113	0.032	0.032	0.007	0.034	0.032
		$\sigma^2$	0.004	0.865	0.848	0.035	1.524	1.486
	Exponential	$\gamma_0$	0.006	0.189	0.187	0.010	0.187	0.190
		$\gamma_1$	0.006	0.285	0.272	0.013	0.269	0.261
		$\rho$	0.193	0.034	0.032	0.023	0.034	0.032
		$\sigma^2$	0.025	0.931	0.879	0.003	1.655	1.561
$\mu_{ij}(\theta) = (2.25)$ $V_{ij}(\theta) =  (2.25) $	Normal	$\gamma_0$	0.023	0.194	0.189	0.013	0.193	0.191
		$\gamma_1$	0.035	0.278	0.274	0.006	0.263	0.264
		$\rho$	0.098	0.033	0.032	0.036	0.035	0.032
		$\sigma^2$	0.008	0.001	0.001	0.031	0.002	0.002
	Exponential	$\gamma_0$	0.007	0.198	0.189	0.007	0.196	0.189
		$\gamma_1$	0.010	0.295	0.275	0.012	0.272	0.264
		$\rho$	0.105	0.034	0.032	0.136	0.035	0.032
		$\sigma^2$	0.015	0.001	0.001	0.003	0.002	0.002

shown), indicating that  $\rho$  is often underestimated. However, as expected, this bias drops dramatically with an increase in sample size. Comparing results for  $E[C_i^*] = 125$  versus  $E[C_i^*] = 225$ , we further observe that a decrease in the number of complete times generally leads to substantially increased standard errors, as might be expected. Interestingly, with a mean censoring time of 125, the use

Table 2.2: Simulation Results for  $n=50$ ,  $C_i \sim N(125, 0)$ 

Expected number of events $\doteq 3.9$			True $F_0$					
Model	Imputed $F_0$	Parameter	Normal			Exponential		
			rBias	ESE	ASE	rBias	ESE	ASE
$\mu_{ij}(\theta) = (2.25)$ $V_{ij}(\theta) = (2.24)$	Normal	$\gamma_0$	0.075	0.469	0.463	0.012	0.455	0.458
		$\gamma_1$	0.152	0.713	0.701	0.011	0.664	0.687
		$\rho$	0.069	0.058	0.057	0.066	0.062	0.061
		$\sigma^2$	0.004	1.206	1.236	0.037	2.173	1.985
	Exponential	$\gamma_0$	0.025	0.489	0.475	0.017	0.489	0.474
		$\gamma_1$	0.078	0.736	0.706	0.005	0.708	0.710
		$\rho$	0.104	0.060	0.058	0.024	0.062	0.061
		$\sigma^2$	0.015	1.331	1.274	0.006	2.247	2.136
$\mu_{ij}(\theta) = (2.25)$ $V_{ij}(\theta) =  (2.25) $	Normal	$\gamma_0$	0.044	0.475	0.466	0.032	0.471	0.460
		$\gamma_1$	0.094	0.727	0.702	0.067	0.702	0.688
		$\rho$	0.092	0.064	0.059	0.182	0.067	0.061
		$\sigma^2$	0.014	0.001	0.001	0.047	0.002	0.002
	Exponential	$\gamma_0$	0.028	0.472	0.470	0.031	0.463	0.474
		$\gamma_1$	0.060	0.720	0.698	0.046	0.698	0.701
		$\rho$	0.172	0.064	0.058	0.047	0.068	0.063
		$\sigma^2$	0.008	0.001	0.001	0.015	0.003	0.002

of a non-constant censoring time often leads to lower standard errors in comparison with a fixed censoring time; however, with a mean censoring time of 225, the situation is reversed. The exact reasons for this change in behavior are unclear, though may have something to do with the fact that the BMI variable changes from a decreasing function at 195 days, a time that lies in between these

two censoring times.

Finally, we remark that Table 2.1 demonstrates a substantial increase in the standard error of  $\hat{\gamma}_1$  in comparison with Table II in Murphy et al. (1995). We repeated the simulation corresponding to Table 2.1, replacing the analytical computation (2.11) with the Monte Carlo approximation described in Section 2.2.4 using  $B = 10000$ . The empirical and estimated asymptotic standard errors were very similar to those in Table 2.1; hence, such a large discrepancy in the estimated standard error of  $\hat{\gamma}_1$  almost certainly reflects the use of an incorrect standard error formula by Murphy et al. (1995), as discussed earlier in Section 2.2.4.

## 2.4 Data Analysis: Recurrent asthma in children

In this section, we use our methodology to reanalyze patterns of recurrent asthma events occurring in young children. Briefly, 232 children aged 6 months at high risk of experiencing an asthmatic event, but who have not yet done so, were randomized to receive either drug or placebo and then followed for up to 18 months. The data are available from the data archive at <http://blackwellpublishing.com/rss>. Duchateau et al. (2003, Table 1) analyze these data using various frailty models and time scales for recurrent event counting processes and find a statistically significant treatment effect. Such a difference seems apparent in Figure 2.1, which respectively summarizes the number of asthmatic events experienced in the drug and placebo groups over the course of the trial.

The gap times of particular interest in this study are the “asthma free” periods, namely (i) the time elapsed between randomization and the beginning of

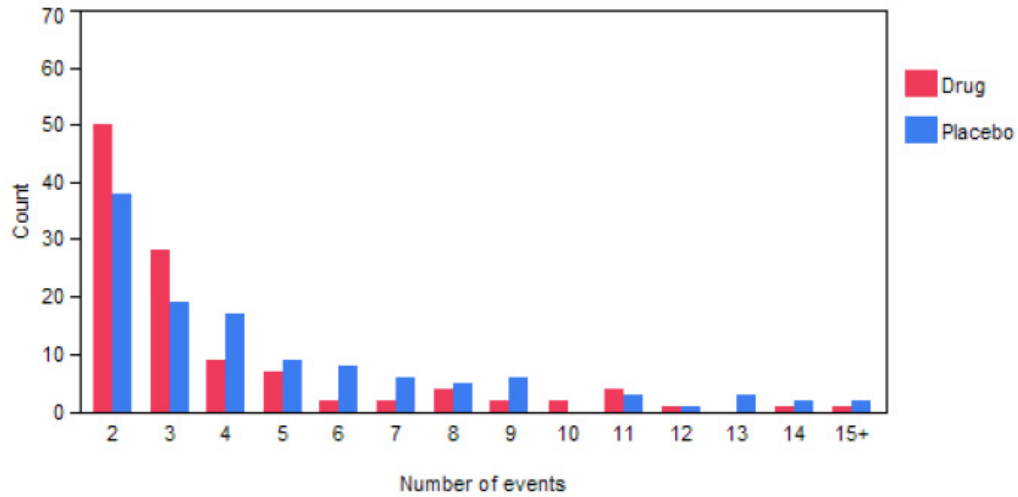


Figure 2.1: Distribution of numbers of recurrent events for drug and control subjects

the first asthmatic episode; and, (ii) the time elapsed between the end of one asthmatic episode and the beginning of the next attack. A complicating factor in this analysis is the fact that a child is not considered to be at risk for another asthmatic event until the current episode ends, with such periods possibly lasting several days.

However, with a median length of 4 days and 95% of the asthmatic events having durations less than 20 days, such times are generally short in comparison with the asthma-free periods (median of 39 days, with 95% of the times less than 430 days). For the purposes of this analysis, we therefore focus on the asthma-free gap times, accounting for the potential impact of asthmatic episodes through covariate adjustment. A secondary analysis, conducted by redefining the gap times of interest as the time elapsed between the start of each asthma-free period, led to the same qualitative and similar quantitative conclusions (results not shown).

In addition to evaluating the treatment effect using various frailty models, Duchateau et al. (2003) comment that there is interest in “the evolution of the asthma recurrent event rate over time,” “how the appearance of an event influences the event rate,” and “how the asthma event rate changes with age.” As suggested in Aalen et al. (2004), heterogeneity (i.e., frailty) can sometimes be accounted for using covariate information that changes over the course of observation. Thus, in the context of modeling gap time data as proposed here, one might investigate how the average length of the current asthma-free episode depends on treatment, the occurrence and length of prior asthmatic and asthma-free episodes, and child age. Due to a significant right skew in the complete gap times, we model the conditional mean of  $Y_{ij} = \log X_{ij}$  (i.e.,  $\mu_{ij}(\theta)$ ), as in (2.5). Two models, described later, are fit assuming  $\mu_{ij}(\theta)$  depends on some function of the covariates  $\bar{L}_{ij} = (D_i, \bar{F}_{ij}, \bar{N}_{ij}, \bar{R}_{ij}, \bar{A}_{ij})^T$ , where  $D_i = I\{\text{child took the drug}\}$ ,  $\bar{F}_{ij} = I\{j > 1\}$  (i.e., “0” for the first event, 1 otherwise),  $\bar{N}_{ij}$  is the length of the most recent asthmatic episode ( $\bar{N}_{i1} = 0$ ),  $\bar{R}_{ij}$  is the length of the most recent asthma-free episode ( $\bar{R}_{i1} = 0$ ), and  $\bar{A}_{ij}$  is the age of the child (in days) at the beginning of the  $j^{\text{th}}$  asthma-free period. To account for the possibility that the  $Y'_{ij}$ s might be heavy tailed, the cumulative distribution function  $F_0(\cdot)$  is chosen as a standardized  $t$  with 3 degrees of freedom. The use of  $F_0(\cdot) = \Phi(\cdot)$  results in no qualitative and minimal quantitative changes (results not shown). The results, reported in Tables 2.3 and 2.4, use  $V_{ij}(\theta) = 1$ ; in general, selecting  $V_{ij}(\theta) = |\mu_{ij}(\theta)|$  leads to nearly identical answers for the regression coefficients  $\theta$  but rather different estimates of  $\sigma^2$ . The reported standard errors (in brackets) are based on (2.18).

Define  $\bar{N}_{ij}^{(b)} = I\{\bar{N}_{ij} \geq 5 \text{ days}\}$  and  $\bar{R}_{ij}^{(b)} = I\{\bar{R}_{ij} \geq 40 \text{ days}\}$  to respectively represent the lengths of the most recent asthmatic and asthma-free episodes

as being above and below the sample median values. Also, let  $\bar{A}_{ij}^{(1)} = I\{\bar{A}_{ij} \in [366, 547] \text{ days}\}$  and  $\bar{A}_{ij}^{(2)} = I\{\bar{A}_{ij} \geq 548 \text{ days}\}$  denote indicators of a child's age to be 1-1.5 years or greater than 1.5 years. Then, the two mean models summarized in Table 2.3 are described below:

1.  $\mu_{ij}(\theta) = \theta_0 + \theta_2 \bar{F}_{ij} + \theta_3 \bar{N}_{ij}^{(b)} + \theta_5 \bar{R}_{ij}^{(b)} + \theta_6 \bar{A}_{ij}^{(1)} + \theta_7 \bar{A}_{ij}^{(2)}$
2.  $\mu_{ij}(\theta) = \theta_0 + \theta_1 D_i + \theta_2 \bar{F}_{ij} + \theta_3 \bar{N}_{ij}^{(b)} + \theta_4 (D_i \times \bar{N}_{ij}^{(b)}) + \theta_5 \bar{R}_{ij}^{(b)} + \theta_6 \bar{A}_{ij}^{(1)} + \theta_7 \bar{A}_{ij}^{(2)}$

In Model 2, the intercept  $\theta_0$  has a useful interpretation as the average of the logarithm of the first gap time for subjects taking placebo. Notably, Model 1 is not fit directly; rather, this model is fit separately by treatment group in order to help illuminate important interactions with treatment, resulting in Models 1a (drug group) and 1b (control group). The intercepts in Models 1a and 1b respectively represent the mean log first gap times for subjects taking drug and placebo. Model 2 assumes a parametric form for the interaction between  $D_i$  and  $\bar{N}_{ij}^{(b)}$  and is therefore a more restricted version of Models 1a and 1b.

Table 2.3: Estimated regression coefficients and standard errors, Models 1 and 2

	$\theta_0$	$\theta_1$	$\theta_2$	$\theta_3$	$\theta_4$	$\theta_5$	$\theta_6$	$\theta_7$
Model 1a	4.874 (0.163)		-0.758 (0.281)	-0.012 (0.155)		0.846 (0.188)	-0.204 (0.180)	-0.846 (0.234)
Model 1b	4.404 (0.194)		-0.771 (0.382)	0.312 (0.139)		0.748 (0.287)	-0.434 (0.171)	-0.230 (0.144)
Model 2	4.420 (0.161)	0.437 (0.175)	-0.747 (0.257)	0.313 (0.134)	-0.449 (0.202)	0.766 (0.185)	-0.325 (0.125)	-0.457 (0.143)

Respectively, the results for  $\theta_0$  under Model 1 and  $\theta_1$  under Model 2 summa-

rized in Table 2.3 indicate the presence of a treatment effect on the initial asthma-free episode, the average gap time being larger in the drug group. Through  $\theta_2$ , both models also indicate that the mean gap time decreases after the occurrence of the first asthmatic episode, the magnitude of this effect varying little by treatment status. Qualitatively, these results are consistent with those of the gap-time-intensity models reported in Table 1 of Duchateau et al. (2003, Models 2 & 3), where statistically significant differences by treatment and between the effects of the first and subsequent events are reported.

Together, Models 1a and 1b further suggest that the mean gap times for the drug and control groups tend to be considerably closer to each other for  $\bar{N}_{ij}^{(b)} = 1$  (i.e., among children whose most recent asthmatic attack is on the longer side) than for  $\bar{N}_{ij}^{(b)} = 0$ . These patterns are present whether or not the effects corresponding to  $\theta_5 - \theta_7$  are included in the model (results not shown). Models 1a and 1b further suggest that a longer previous asthma free episode tends to result in a longer current asthma free episode, regardless of treatment. We also observe that the average length of asthma-free episodes tends to decrease with increasing age, with the drug having a relatively protective effect in younger children (i.e., less than 1.5 years old) that may start to wear off with increasing age.

Models 1a and 1b also demonstrate the existence of a modest level of interaction between certain patient history variables and treatment. Specifically, there are noticeable changes in  $\theta_3$  (length of most recent asthmatic episode) and  $\theta_7$  (children older than 1.5 years) and more minor changes in  $\theta_5$  (length of most recent asthma-free episode) and  $\theta_6$  (children aged 1-1.5 years). However, with the exception of  $\theta_3$ , these effects only exhibit changes in magnitude, not in direc-

tion. Therefore, Model 2 can be expected to provide a parsimonious description of the trends exhibited in Models 1a and 1b, as well as a more direct evaluation of the treatment effect. In fact, under Model 2, we observe that the treatment effect and its interaction with  $\bar{N}_{ij}^{(b)}$  are both statistically significant ( $p = 0.012$  and  $p = 0.026$ , respectively).

Model 3, summarized in Table 2.4, is specified similarly to Model 2, except that continuous versions of certain covariates are used in order to investigate the impact of discretizing covariates. Let  $\bar{N}_{ij}^{(log)} = 0$  if  $j = 1$  and  $\bar{N}_{ij}^{(log)} = \log \bar{N}_{ij}$  for  $j > 1$ ; define  $\bar{R}_{ij}^{(log)}$  similarly. Also, let  $\bar{A}_{ij}^{(log)} = \log \bar{A}_{ij} - \log 182$  be the log age of a child, centered at its minimum value of 182 days. Then, the mean function for Model 3 is the following:

$$\mu_{ij}(\theta) = \theta_0 + \theta_1 D_i + \theta_2 \bar{F}_{ij} + \theta_3 \bar{N}_{ij}^{(log)} + \theta_4 D_i \bar{N}_{ij}^{(log)} + \theta_5 \bar{R}_{ij}^{(log)} + \theta_6 \bar{A}_{ij}^{(log)}.$$

The results of fitting this model reflect the same trends observed in Model 2.

Table 2.4: Estimated regression coefficients and standard errors, Model 3

	$\theta_0$	$\theta_1$	$\theta_2$	$\theta'_3$	$\theta'_4$	$\theta'_5$	$\theta'_6$
Model 3	4.376 (0.162)	0.529 (0.167)	-1.263 (0.349)	0.121 (0.071)	-0.258 (0.107)	0.369 (0.063)	-0.778 (0.136)

Residual analysis for GEE models with longitudinal data, particularly in the presence of missing data, is not a well-developed field. In the current setting, one might consider using

$$\widehat{W}_{ij} = \frac{Y_{ij} - \mu_{ij}(\widehat{\theta})}{\widehat{\sigma} V_{ij}(\widehat{\theta})}, \quad j = 1, \dots, N_i, \quad i = 1, \dots, n,$$

that is, the estimated “standardized residuals” derived from the complete gap time information. Figure 2.2 provides histograms of these quantities respec-



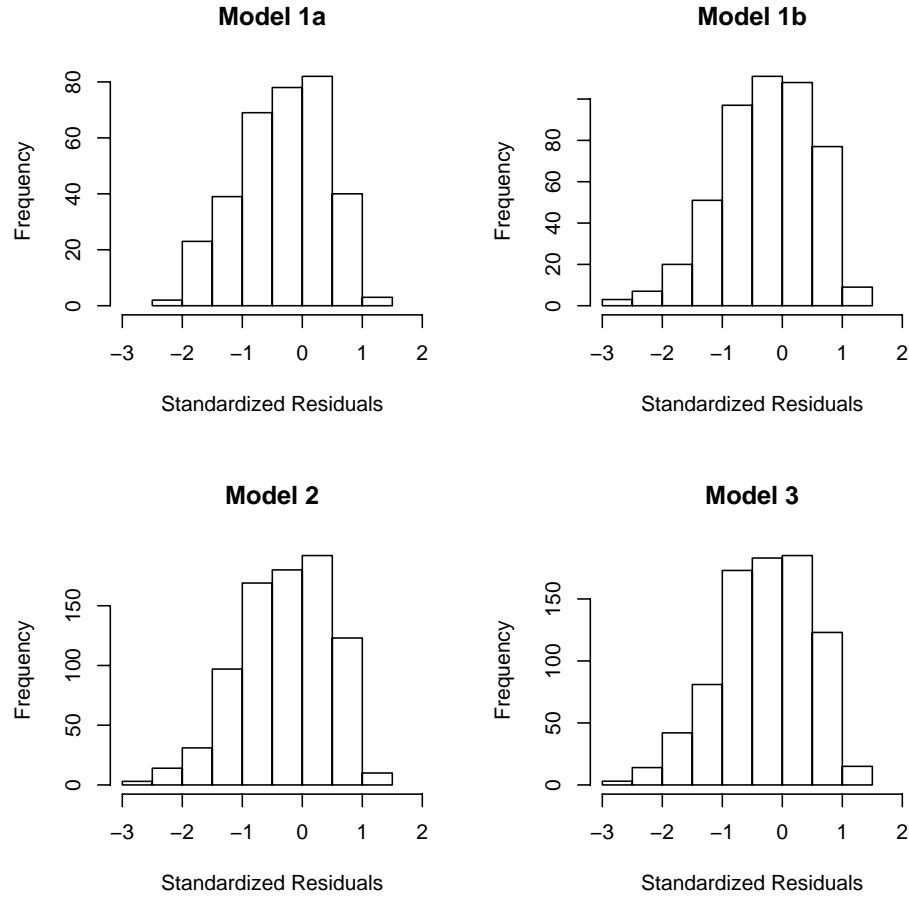


Figure 2.2: Standardized complete gap time residuals for Models 1-3

tively obtained under Models 1-3. While such plots cannot be used to validate that the individual mean and variance functions have been well-specified, the lack of unusually large standardized residuals is an indication that the model has done a reasonable overall job in describing the observed gap time data. However, care must be exercised in the interpretation of such plots due to the presence of correlation between residuals and also the fact that  $\sum_{i=1}^n \sum_{j=1}^{N_i} \widehat{W}_{ij} \neq 0$ . The last result is a consequence of using residuals derived only from complete gap time information, and likely explains the mild left skew observed in each plot.

## 2.5 Discussion

The use of GEE-type methods for analyzing recurrent event counting processes is now common. In contrast, the use of such methods for directly analyzing gap time data has not been systematically investigated. The methodology proposed here extends and corrects methodology originally proposed in Murphy et al. (1995) for the purposes of analyzing menstrual cycle data. The result is a simple yet flexible class of models for analyzing gap time data. An especially attractive feature is the ability to specify rich models through mean and variance structures, leading to direct interpretability of regression effects on the gap times.

The main limitation of the proposed methodology is the reliance of (2.14) and (2.15) on the parametric model specified by (2.11) and (2.10). This parametric assumption is used for the sole purpose of dealing with censored gap time information and our simulation results also demonstrate a degree of robustness to the misspecification of  $F_0(\cdot)$ . However, for reasons explained earlier, both consistency and asymptotic normality of  $\widehat{\eta}_n$  do rely on the correct specification of the imputation model (2.10). Consequently, it is worthwhile to consider estimating  $\eta$  using alternative methods for handling the censored gap times.

For example, one might try and estimate the required conditional moments for the censored gap times under (2.10) without imposing a specific parametric model in (2.11). Let  $M > 0$  be an arbitrarily large fixed integer and define for  $i = 1, \dots, n$  and  $j = 1, \dots, M$  the probabilities

$$\pi_{ij} = P\{C_i \geq S_{ij} | \mathcal{H}_{ij} \cup S_{ij}\}. \quad (2.26)$$

Assume that  $\pi_{ij} \geq \epsilon > 0$  for  $i, j \geq 1$  and, in addition, that  $E[W_{ij}^r(\eta_0) | W_{ij}(\eta_0) > w] =$

$K_r(w)$ ,  $r = 1, 2$  and  $i, j \geq 1$ . Under mild regularity conditions,

$$\widehat{K}_r(w, \eta_0) = \frac{\sum_{i=1}^n \sum_{j=1}^M \frac{I\{C_i \geq S_{ij}\}}{\pi_{ij}} W_{ij}^r(\eta_0) I\{W_{ij}(\eta_0) > w\}}{\sum_{i=1}^n \sum_{j=1}^M \frac{I\{C_i \geq S_{ij}\}}{\pi_{ij}} I\{W_{ij}(\eta_0) > w\}}$$

is a pointwise consistent estimator of  $K_r(w)$  for  $r = 1, 2$ . This estimator avoids the need to use a parametric specification for  $F_0(\cdot)$  in (2.11) at the price of assuming that  $E[W_{ij}^r(\eta_0) | W_{ij}(\eta_0) > w] = K_r(w)$ ,  $r = 1, 2$  for all standardized gap times. This latter assumption is stronger than that made in connection with (2.10), which is only imposed on the last incomplete gap time on each subject. In practice, use of this inverse-probability-of-censoring-weighted (IPCW) estimator also requires estimating  $\eta_0$  and the censoring probabilities in (2.26). In the special case where censoring is completely independent of the event process and covariates, it is possible to consistently estimate the  $\pi_{ij}$  via  $\widehat{\pi}_{ij} = 1 - \widehat{G}(S_{ij+})$ , where  $\widehat{G}(\cdot)$  is the empirical CDF of the censoring times  $C_1, \dots, C_n$ . More generally, a model for the censoring mechanism must be imposed; if this model is misspecified, bias can be expected for reasons analogous to the misspecification of  $F_0(\cdot)$ . For these and other reasons (e.g., the need for uniform asymptotic results, proper methods of variance estimation), we have not investigated the utility of this estimator any further.

An alternative use of IPCW estimation is to construct direct analogs of (2.14) and (2.15). Specifically, one might proceed by estimating  $\eta$  using

$$\frac{1}{n} \sum_{i=1}^n \left\{ \sum_{j=1}^M \frac{I\{S_{ij} \leq C_i\}}{\pi_{ij}} f_{ij}(\theta) W_{ij}(\eta) \right\} \quad \text{and} \quad \frac{1}{n} \sum_{i=1}^n \left\{ \sum_{j=1}^M \frac{I\{S_{ij} \leq C_i\}}{\pi_{ij}} b_{ij}(\eta) (W_{ij}^2(\eta) - 1) \right\}.$$

While such an approach obviates the need to impose assumptions (2.11) and (2.10), the avoidance of bias continues to require that one be able to correctly model and consistently estimate the censoring probabilities (2.26). In addition, because the information on censored gap times is no longer utilized directly,

the efficiency of such an approach may suffer. This efficiency loss may be offset by including information on the censored gap times via augmented estimating equations (e.g., Rotnitzky et al., 1998; Scharfstein et al., 1999; Bang and Robins, 2005).

Finally, the methods developed in this chapter can in principle be extended to multivariate recurrent event processes arising either as a result of having clustered data or due to the presence of multiple recurrent event outcomes on each subject. However, if the dependence structure between processes or especially between censored multivariate gap times must be modeled, the robustness of the present approach is likely to suffer and it may be advantageous to use a proper extension of the IPCW-type estimation scheme described above.

CHAPTER 3

MARGINAL AND CONDITIONAL ESTIMATING EQUATIONS FOR  
MULTIVARIATE LONGITUDINAL DATA SUBJECT TO CENSORING

### 3.1 Introduction

Longitudinal observations obtained at discrete times are often the only means available for describing the evolution of a (multivariate) response of interest. When these discrete times, often called visit times, are themselves informative of the response (e.g. patients undergoing more frequent testing tend to also be ones who are unhealthy) the analysis is complicated, and a likelihood approach can lead to biased estimates of the parameters driving the underlying longitudinal process of interest. Typically interest is in this underlying process, rather than in the observed process.

Estimating parameters in an underlying continuous process based on discrete sampling has long been a topic of interest in the financial literature, in which this estimation is integral to theoretical asset pricing. An extensive list of references is available in this area, e.g. Kessler and Sørensen (1999); Kessler (2000); Bibby et al. (2004), with a detailed review by Aït-Sahalia (2005). These authors assume the underlying process,  $X(t)$ , follows a diffusion with the stochastic differential equation  $dX(t) = b(X(t), \theta)dt + \sigma(X(t), \theta)dW(t)$ , where  $b$  and  $\sigma$  are known functions,  $W(t)$  is a Brownian motion, and  $\theta$  is an unknown parameter vector. They use unbiased estimating equations (sometimes based on the eigenfunctions of the generator of the diffusion) to estimate the true  $\theta$ .

The extensive theoretical aspect of these works is not the focus of this thesis,

but thinking of the longitudinal process as a multidimensional diffusion process - or, more generally, modeling the mean and variance of its transition density - is an interesting idea. However, the above references, and the financial literature in general, only consider univariate diffusions unfolding over a long time period on one subject, rather than many subjects, possibly with subject-specific time-fixed covariates to incorporate into the model. Furthermore, the aforementioned papers do not consider informative observation times.

Aït-Sahalia and Mykland (2003) deal with informative visitation times, where the time interval,  $\Delta$ , between visits is chosen randomly at the beginning of the interval, and may depend only on the length of the previous interval and on the location of the diffusion,  $X$ ; the conditional methodology proposed in this chapter will require a similar assumption, which is outlined in Section 3.2.3. Their method involves breaking the joint likelihood of  $(X, \Delta)$  into two pieces, the second of which relates to the visitation process and does not depend on  $\theta$ , the unknown parameter vector of interest. Hence, they obtain an estimate of  $\theta$  by maximizing the first piece, which is based on the transition density of the observed diffusion.

Moving from the diffusion process setting to the longitudinal data setting, Lipsitz et al. (2002) also break the joint likelihood into a piece for the response process and a piece for the visitation process, only the first of which depends on the regression parameters of interest. Similarly to Aït-Sahalia and Mykland (2003), they accomplish this with an ignorability-type assumption regarding the visitation process, which simplifies the analysis. Lin and Ying (2001), Lin et al. (2004), and Bůžková and Lumley (2009) use progressively weaker assumptions on the visitation process to estimate regression parameters in a univariate

marginal response model by way of estimating equations. The latter two papers analyze HUD-VASH homelessness data, which we will reanalyze in Section 3.4.1 from a different standpoint.

Through the use of a latent variable in both the response and visitation models, recent papers by Sun et al. (2007) and Liang et al. (2009) generalize the setting of Bůžková and Lumley (2009) by allowing the response process and the visitation process to be dependent even when conditioned on the observed process histories. While this generalization is meaningful, these two papers both require the visit process intensity to depend only on the time-fixed latent variable and time-fixed covariates. This precludes observation times that depend on the response process or on any time-dependent covariates. Sun and Tong (2009) make use of the more restrictive framework in Sun et al. (2007) to also include a weaker non-informative censoring assumption, specifically by adding a latent variable to the conditioning statement. Sun et al. (2005) allow the conditional response process to depend on a function of the past visit process, as well as the covariates, as another way of introducing dependence between the two processes.

The approach to parameter estimation in the longitudinal setting that is presented in this chapter is different from the current literature, e.g. Bůžková and Lumley (2009), for four main reasons. First, previous observations of our response process of interest may serve as predictors in our estimating equations, making the EEs potentially conditional. This is especially useful when the ultimate goal may be to use the estimated subject-specific trajectories to predict future trajectories or to predict an associated time-to-event. When previous responses are not included in the conditioning statements for the mean and vari-

ance, the estimating equations become “marginal”; the second difference between our work and that of the aforementioned authors is that our marginal estimating equations allow the covariates to be related to the mean response in an arbitrary fashion, whereas Lin and Ying (2001) and Bůžková and Lumley (2009) assume that the marginal mean function is linear in the time-dependent covariates, and Lin et al. (2004) model the response as a function of only time-fixed covariates. Third, we estimate variance parameters, which generalizes GEE for variance parameters (Prentice, 1988; Zhao and Prentice, 1990; Prentice and Zhao, 1991) by considering informative observation times, and fourth, we permit a multivariate response, requiring estimation of correlation between its different dimensions.

The rest of the chapter proceeds as follows. Section 3.2.1 introduces the notation, while Sections 3.2.2 and 3.2.3 focus respectively on marginal estimating equations and conditional estimating equations. Section 3.2.3 contrasts the assumptions required for the conditional model to those required for the marginal model. Both sections provide an estimating equation for parameters indexing the mean and variance of the process of interest. Section 3.3 provides results of simulations, and Sections 3.4.1 and 3.4.2 respectively analyze the univariate HUD-VASH data and the multivariate protein data. Technical details and large sample theory are provided in Appendices E and G.



## 3.2 Methodology

### 3.2.1 Notation

Let the longitudinal response process of interest be  $X(t)$ , and let the available covariates be  $A(t)$ . Note that  $A$  may include time-dependent internal covariates (as defined in Kalbfleisch and Prentice, 2002, Section 6.3.2); it may also include time-fixed covariates measured upon entry into the study. The time-dependent covariates and the response for subject  $i$  are measured at each of subject  $i$ 's visitation (i.e. measurement, observation, or test) times,  $0 \leq t_{i0} < t_{i1} < \dots < t_{iK_i} < C_i < \tau$ , where  $C_i$  is the end of the observation period for subject  $i$  (the incorporation of missing data and failure times will be covered in Chapter 4). We will often utilize the term “visitation” because it is clear what is meant by visitation (even it isn't always technically a “visit”), whereas measurement, for example, could mean the measurement times, or the measurements themselves. Let  $N_i(t) = \sum_{k=1}^{K_i} I(t_{ik} \leq t)$  count the number of visits up to and including time  $t$ ,  $\xi_i(t) = I(C_i \geq t)$  be the at-risk indicator for subject  $i$ , and  $N_i^*(t)$  be the underlying uncensored process with  $N_i(t) = N_i^*(t \wedge C_i)$ . The reason to possibly define  $0 \leq t_{i0}$ , rather than  $0 = t_{i0}$ , is to effectively synchronize the time scales for different subjects, enabling estimation of a cyclical time trend in the visit process (e.g. non-informative measurements taken at 6am every day in a hospital are accounted for by setting time zero for subject  $i$  to be midnight on the day of subject  $i$ 's first measurement).

For subject  $i$ ,  $\bar{X}_i(t)$  and  $\bar{A}_i(t)$  will be used to denote the respective observed histories, up to and including time  $t$ , of the response and covariate processes, with lack of the “bar” denoting the (possibly unobserved) value at time  $t$  itself.

For convenience, the times of the observations for subject  $i$  will be included implicitly in  $\bar{X}_i(t)$  and  $\bar{A}_i(t)$ .

The following two subsections, 3.2.2 and 3.2.3, respectively present the marginal and conditional approaches to modeling longitudinal data.

### 3.2.2 Marginal model

#### Assumptions

The main model assumptions of this subsection are

$$E(X_i(t)|H_i^{m,m}(t)) = \mu(h_i^{m,m}(t), \theta_0^m) \equiv \mu_i(t), \quad (3.1)$$

$$Var(X_i(t)|H_i^{m,m}(t)) = \Sigma(h_i^{m,m}(t), \theta_0^m) \equiv \Sigma_i(t), \quad (3.2)$$

for some known functions  $\mu$  and  $\Sigma$  and unknown parameter vector  $\theta_0^m$ , where  $\{H_i^{m,m}(t)\} = \{A_i(t)\}$  and  $h_i^{m,m}(t)$  is a realization of  $\{H_i^{m,m}(t)\}$ . The dependence on  $\theta$  has been suppressed in (3.1) and (3.2) for parsimony in the notation. To explain the notation, the first  $m$  in the superscript stands for moment because  $\{H_i^{m,m}(t)\}$  is the set conditioned on to model the mean and variance, and the second  $m$  stands for marginal;  $\{H_i^{v,m}(t)\}$  will denote the set conditioned on to model the visitation times, and  $\{H_i^{m,c}(t)\}$  will denote the set conditioned on to model the conditional moments.

Marginal modeling of  $X_i(t)$  conditions on  $A_i(t)$ , but not on the observed history of the response,  $\bar{X}_i(t^-)$ . Also, one must be careful when explicitly conditioning on  $A_i(t-k)$  for  $k > 0$  (in the informative visitation setting) because only values of  $k$  such that  $dN_i(t-k) = 1$  would be possible, creating an implicit term in the

mean model indicating a lack of visitation in certain intervals. When  $\{H_i^{v,m}(t)\}$  includes  $X_i(t)$ , this leads to a biased estimating equation in general; this issue is discussed more fully in Section 3.2.3 and in Appendix E. Rather than explicitly conditioning on  $A_i(t - k)$  for  $k > 0$ , the covariate history can safely be included in  $A_i(t)$  in summary form, e.g. an average of all previously observed covariates, but such a model would implicitly include aspects of visitation history, which can in general depend on the history of the response, leading to a model that is no longer fully marginal.

When inferences about the population average are the ultimate goal, marginal modeling is appropriate. For example, in the HUD-VASH data analyzed by Lin et al. (2004) and Bůžková and Lumley (2009), interest lies in comparing the effect of three different types of intervention for homelessness on a population level, so the authors take a univariate marginal approach to model only the first moment. In general, the methods in these papers specify that the covariates be related linearly to the mean response; we relax this assumption in (3.1).

Marginal likelihood or GEE (Liang and Zeger, 1986) could be used to estimate  $\theta$  from the discrete observations. However in many settings, including the one in this chapter, the discrete observation times themselves may be informative about the process  $X_i$ , and in those cases the maximizer of the likelihood or the solution to the GEE is biased for  $\theta_0^m$ , i.e. for the parameter that drives the *underlying* response process as opposed to the observed response process. One solution to this issue, while maintaining relatively weak assumptions on the visit process, is to use a weighted estimating equation with weights inversely proportional to the visitation intensity. This in effect creates a pseudo-population

in which the visit process and the response process are no longer associated (Lin et al., 2004). Since we are solving for variance parameters as well as mean parameters, the weighted estimation equation proposed here is actually a generalization of GEE for variance parameters (Prentice, 1988). The method is a special case of GEE2 (Zhao and Prentice, 1990, Prentice and Zhao, 1991), but since the variance is not forced to be a function of the mean, the joint estimation of mean and variance parameters that defines GEE2 is not possible.

We assume:

$$E\left(X_i(t)^{\otimes j} | H_i^{m,m}(t), C_i \geq t\right) = E\left(X_i(t)^{\otimes j} | H_i^{m,m}(t)\right), \quad (3.3)$$

$$E\left(dN_i^*(t) | H_i^{m,m}(t), H_i^{v,m}(t), X_i(t), C_i \geq t\right) = E\left(dN_i^*(t) | H_i^{v,m}(t)\right), \quad (3.4)$$

where  $j = 1, 2$ ,  $x^{\otimes 1} = x$  and  $x^{\otimes 2} = xx^T$  for a vector  $x$ ,  $dN_i^*(t) = N_i^*(t^- + dt) - N_i^*(t^-)$ , and  $\{H_i^{v,m}(t)\} = \{\bar{X}_i(t^-), X_i(t), \bar{A}_i(t^-), A_i(t)\}$ . Assumption (3.3) is an independent censoring assumption and (3.4) is termed by Bůžková and Lumley (2009) as an independent sampling assumption. Most importantly the latter allows for visitation times dependent on responses. To see how assumption (3.4) compares with the conditional model of Section 3.2.3 and with other relevant literature, see Appendix E.

### Marginal estimating equation

The estimating equation below will solve for the complete parameter vector  $\theta$ , but it's notationally helpful to define  $\theta = (\beta^T, \alpha^T)^T$  where  $\beta$  and  $\alpha$  are the vectors of mean and variance parameters respectively. Let  $s_i^*(t)$  be the upper triangle of  $s_i(t) = (X_i(t) - \mu_i(t))(X_i(t) - \mu_i(t))^T$  written as a vector, and  $\Sigma_i^*(t)$  be the upper triangle of  $\Sigma_i(t)$  in vector form. The proposed marginal estimating equation is

then:

$$U^m(\theta, \hat{\gamma}) = \sum_{i=1}^n \int_0^\tau \left( \frac{\partial \mu_i(t)}{\partial \beta^T} V_{1i}^{-1}(t) (X_i(t) - \mu_i(t)) \right) f(h_i^{v,m}(t)) w_i^v(t, \hat{\gamma}) dN_i(t), \quad (3.5)$$

where  $\tau$  is such that  $P(C_i > \tau) > 0, \forall i$ ,  $w_i^v(t, \gamma) = \exp(-\gamma^T h_i^{v,m}(t))$ ,  $h_i^{v,m}(t)$  and  $h_i^{v,m}(t)$  are vectors of information available in  $\{H_i^{m,m}(t)\}$  and  $\{H_i^{v,m}(t)\}$  respectively,  $V_{1i}(t)$  and  $V_{2i}(t)$  are “working” estimates of the variances of  $X_i(t)$  and  $s_i^*(t)$  respectively,  $f$  is any known function mapping a finite-dimensional vector to a scalar, and

$$E(dN_i^*(t) | H_i^{v,m}(t)) = \exp(\gamma_0^T h_i^{v,m}(t)) d\Lambda_0(t), \quad (3.6)$$

where  $\Lambda_0(t)$  is a non-decreasing function of time. Estimation of  $\theta_0^m$  is a two step process because before solving (3.5),  $\gamma_0$  in (3.6) must be estimated. One way to do this is by solving

$$U^*(\gamma) = \sum_{i=1}^n \int_0^\tau (h_i^{v,m}(t) - \bar{h}_i^{v,m}(t, \gamma)) dN_i(t), \quad (3.7)$$

where

$$\bar{h}_i^{v,m}(t, \gamma) = \sum_{i=1}^n h_i^{v,m}(t) \frac{\xi_i(t) \exp(\gamma^T h_i^{v,m}(t))}{\sum_{j=1}^n \xi_j(t) \exp(\gamma^T h_j^{v,m}(t))}.$$

As in Buřková and Lumley (2009) we can accommodate a visit hazard that may not be purely continuous. Note that no smooth hazard rate needs to be estimated in order to solve (3.5), and  $\hat{\Lambda}_0(t)$  does not show up in any calculations except the large sample variance of  $\hat{\theta}^m$ . We leave  $\Sigma_i(t)$  completely unspecified but it would be possible to lessen the computational burden by assuming exchangeability, for example.

There are two possible layers of dependence within subject for an estimating equation with a multidimensional response: dependence between the dimensions within visit (which (3.5) allows for, and in fact estimates) and dependence

between visits. Thinking of the variance for each subject in block form, where the diagonal imposes dependence within visit and the off-diagonal imposes dependence between visits, (3.5) is assuming a block diagonal form. This block diagonal form is needed to ensure unbiasedness of the resulting parameter estimates given a weak visit process assumption like (3.4). If the variance between visits were to be estimated, unbiasedness would require the response to be conditionally independent of future covariates given the current covariate, a condition outlined in Pepe and Anderson (1994). This requirement arises from the fact that a non-diagonal variance creates contributions to the sum in (3.5) that depend on the response minus its mean, multiplied by future covariates. To be more precise, (3.5) would have contributions like  $(X(t) - \mu(t))f(A(s+t))$  for some function  $f$  and  $s > 0$ . Such contributions only have mean zero if  $X(t)$  is conditionally independent of  $A(s+t)$  given  $A(t)$ . With internal covariates in particular, this is a strong assumption.

**Theorem 3.2.1.** Given (3.1)-(3.4), (3.6), and regularity conditions (G1)-(G11) in Appendix G.1, and assuming that  $\hat{\gamma}$ , the solution to (3.7), is consistent for  $\gamma_0$ , the  $\hat{\theta}^m$  that solves  $0 = U^m(\hat{\theta}^m, \hat{\gamma})$  is a consistent and asymptotically normal estimator of  $\theta_0^m$ , the true parameter.

Assuming  $\hat{\gamma}$  is consistent for  $\gamma_0$  is usually a necessary assumption (e.g. in Lin et al. (2004) and Bůžková and Lumley (2009)) whenever a Cox model with intermittently observed time-dependent covariates is in question. However, consistent estimation of the true  $\gamma_0$  is unlikely in truth; the reader is referred to Andersen and Liestøl (2003) for a discussion of the attenuation in  $\gamma$  and some strategies for bias reduction. The proof of Theorem 3.2.1 is provided in Appendix G.2, along with an estimate of the asymptotic variance of  $\hat{\theta}^m$ . Note that the large sample theory in this thesis is all done as the number of subjects, as

opposed to the number of visits, goes to infinity.

### 3.2.3 Conditional model

This section introduces the conditional estimating equation (CEE) and its required modeling assumptions. Throughout this section, we will contrast the CEE's model and assumptions, both positively and negatively, with the marginal model from Section 3.2.2.

#### Assumptions

The main model assumptions of this subsection are

$$E\left(X_i(t)|H_i^{m,c}(t)\right) = \mu\left(h_i^{m,c}(t), \theta_0^c\right) \equiv \mu_i(t), \quad (3.8)$$

$$Var\left(X_i(t)|H_i^{m,c}(t)\right) = \Sigma\left(h_i^{m,c}(t), \theta_0^c\right) \equiv \Sigma_i(t), \quad (3.9)$$

for some known functions  $\mu$  and  $\Sigma$  and unknown parameter vector  $\theta_0^c$ , where  $\{H_i^{m,c}(t)\} = \{\bar{X}_i(t^-), \bar{A}_i(t^-), A_i(t)\}$  and  $h_i^{m,c}(t)$  is a realization of  $\{H_i^{m,c}(t)\}$ . The  $\mu$  and  $\Sigma$  here are of course not the same as the  $\mu$  and  $\Sigma$  from (3.1) and (3.2), but it's notationally convenient to reuse them. The dependence on  $\theta$  has again been suppressed for parsimony.

Conditional modeling (labeled as “transition modeling” by Diggle et al., 2002) of  $X_i(t)$  conditions on the covariate history,  $\bar{A}_i(t^-)$ , and on the response history,  $\bar{X}_i(t^-)$ . When subject-specific trajectory estimation is desired, conditional models are preferable to marginal models. They are also useful for prediction (of future trajectory or of a time-to-event), but they are more subject to misspecification because of the need to specify a structure for the first two conditional

moments. However, since our ultimate goal is to predict survival (see Chapter 5), we require this subject-specific estimate that the conditional model conveniently provides. The reader should note that  $A_i(t)$  may be included in  $\{H_i^{m,c}(t)\}$  in this chapter and in Chapter 4, but in Chapter 5, where interest will shift to the distribution of  $X_i(t)$  for arbitrary  $t$ , its inclusion will only be possible if  $A_i(t)$  is continuously observed.

The analogs of (3.3) and (3.4) for the conditional model are

$$E\left(X_i(t)^{\otimes j} | H_i^{m,c}(t), C_i \geq t\right) = E\left(X_i(t)^{\otimes j} | H_i^{m,c}(t)\right), \quad (3.10)$$

$$E\left(dN_i^*(t) | H_i^{m,c}(t), H_i^{v,c}(t), X_i(t), C_i \geq t\right) = E\left(dN_i^*(t) | H_i^{v,c}(t)\right), \quad (3.11)$$

where  $j = 1, 2$ ,  $x^{\otimes 1} = x$  and  $x^{\otimes 2} = xx^T$  for a vector  $x$ , and  $\{H_i^{v,c}(t)\} = \{\bar{X}_i(t^-), \bar{A}_i(t^-)\}$ .

We also require

$$E\left(X_i(t)^{\otimes j} | H_i^{m,c}(t), H_i^{v,c}(t), C_i \geq t\right) = E\left(X_i(t)^{\otimes j} | H_i^{m,c}(t), C_i \geq t\right) \quad (3.12)$$

for  $j = 1, 2$ , which has no corresponding marginal model assumption. Assumption (3.12) is required in the conditional model because of the implicit assumption, in the structure of  $\mu_i(t)$ , that no visits have taken place between  $r_i(t)$  and  $t$ , where  $r_i(t) = \max(s : dN_i(s) = 1, s < t)$ ; for more detail, see Appendix E. In practice (3.12) might make little difference because a common goal of a conditional model is to predict the future trajectory of  $X_i(t)$  as accurately as possible, and hence any information in  $\{H_i^{v,c}(t)\}$  should be included in  $\{H_i^{m,c}(t)\}$ . Even if this results in misspecification of (3.8) or (3.9), it is better than not accounting for the information at all. Advantageously, (3.12) implies that no inverse-intensity-weight is required in our CEE, meaning estimation of  $\gamma_0$  is not required. See Appendix G.4 for detail.

Note that (3.10) is generally weaker than (3.3) because  $\{H_i^{m,c}(t)\}$  includes previous covariates and previous responses, whereas  $\{H_i^{m,m}(t)\}$  does not. However,



(3.11) is stronger than (3.4) because the visitation intensity cannot depend on  $X_i(t)$  and  $A_i(t)$ . To see why this is so, note that if  $\{H_i^{v,c}(t)\}$  contained  $X_i(t)$  then (3.12) could not hold anymore. It turns out that to maintain unbiasedness the CEE would then generally require an inverse weight, the calculation of which would not be tractable without jointly modeling the processes  $N_i$  and  $X_i$ , hence destroying the simplicity of (3.8) and (3.9). For the same reasons,  $\{H_i^{v,c}(t)\}$  cannot include any portion of  $A_i(t)$  that is correlated with  $X_i(t)$ . See Appendix E for more detail. To see how assumption (3.11) compares with the marginal model of Section 3.2.2 and with other relevant literature, also see Appendix E.

### Conditional estimating equation

Let  $s_i^*(t)$  be the upper triangle of  $s_i(t) = (X_i(t) - \mu_i(t))(X_i(t) - \mu_i(t))^T$  written as a vector, and  $\Sigma_i^*(t)$  be the upper triangle of  $\Sigma_i(t)$  in vector form. Note that as in Chapter 2,  $\mu_i(t)$  and  $\Sigma_i(t)$  may depend on the response process history; our proposed conditional estimating equation is then:

$$U^c(\theta) = \sum_{i=1}^n \int_0^\tau \left( \begin{array}{c} \frac{\partial \mu_i(t)}{\partial \beta^T} V_{1i}^{-1}(t) (X_i(t) - \mu_i(t)) \\ \frac{\partial \Sigma_i^*(t)}{\partial \alpha^T} V_{2i}^{-1}(t) (s_i^*(t) - \Sigma_i^*(t)) \end{array} \right) f(h_i^{m,c}(t)) dN_i(t), \quad (3.13)$$

where  $\tau$  is such that  $P(C_i > \tau) > 0, \forall i$ ,  $h_i^{m,c}(t)$  is a vector of information available in  $\{H_i^{m,c}(t)\}$ ,  $V_{1i}(t)$  and  $V_{2i}(t)$  are “working” estimates of the conditional variances of  $X_i(t)$  and  $s_i^*(t)$  respectively, and  $f$  is any known function mapping a finite-dimensional vector to a scalar. As in Section 3.2.2, we leave  $\Sigma_i(t)$  completely unspecified but it would be possible to lessen the computational burden by assuming exchangeability, for example.

**Theorem 3.2.2.** Given (3.8)-(3.12), and regularity conditions (H1)-(H5) in Appendix G.3, the  $\hat{\theta}^c$  that solves  $0 = U^c(\hat{\theta}^c)$  is a consistent and asymptotically nor-

mal estimator of  $\theta_0^c$ , the true parameter.

The proof of Theorem 3.2.2 is provided in Appendix G.4, along with an estimate of the asymptotic variance of  $\hat{\theta}^c$ .

### 3.3 Simulations

In this section, the performance of (3.13) will be studied for finite samples. We will focus our modeling on two diffusion processes: one with mean reverting drift and one with a deterministic drift. We focus on the conditional method here for two reasons: (i) it is an innovative contribution of this thesis, and (ii) it will be the focus of Chapter 4 when missingness and failures are introduced. This will allow easy comparison of these simulations with those in Chapter 4. Both the diffusions here have a variance matrix that is constant over  $X$  and  $t$  in order to induce a Gaussian transition density. The model for the transition density does not have to correspond to a stochastic differential equation, but to simulate the data in Chapter 4, particularly the mortality process, the underlying processes need to be updated continuously (while still maintaining a closed form transition density), restricting the cases we can study by simulation. In order to make comparisons between these simulations and the ones of the upcoming chapter, we study those same processes here.

The mean reverting process is often called the Ornstein-Uhlenbeck process. In our simulations, we utilize a subject-specific version of its SDE, specifically:

$$dX_i(t) = B\left(\mu + g(i) - X_i(t)\right)dt + \Sigma^{1/2}dW_i(t),$$

and the corresponding multivariate transition density:

$$(X_i(s+t)|X_i(s), Z_i) \sim N\left(\mu + g(i) + \exp(-Bt)(X_i(s) - \mu - g(i)), \Sigma^\dagger - \exp(-Bt)\Sigma^\dagger \exp(-B^T t)\right),$$

where  $W_i(t)$  are i.i.d. Brownian motions,  $\mu$  is an unknown parameter vector,  $B$  and  $\Sigma$  are unknown non-negative definite matrices of parameters,  $\Sigma^\dagger = (2B)^{-1/2}\Sigma(2B)^{-1/2}$ ,  $g(i) = A_i \times \phi$  is the subject-specific part of the diffusion, where  $\phi$  is an unknown parameter vector, and  $A_i$  is the vector of time-fixed covariates for subject  $i$ . We assume  $B$  is a  $d \times d$  diagonal matrix, which makes  $\exp(B)$  diagonal with entries  $\exp(b_1), \dots, \exp(b_d)$ . The larger  $b_j$  is, the faster the  $j^{th}$  process reverts to its mean. The diagonality assumption is used for computational ease, but also makes some sense in practice: if one or more dimensions of the process have gotten out of line and are in the process of reverting to their means, we don't necessarily expect the rest of the dimensions to have high short term drifts too.

Our deterministic drift has the multivariate transition density:

$$(X_i(s+t)|X_i(s)) \sim N(X_i(s) + (\mu(s+t, \beta) - \mu(s, \beta)), \Sigma \cdot t),$$

where  $\mu(\cdot)$  is a B-spline, and  $\Sigma$  is a matrix of unknown parameters. For parsimony, we assume the B-spline is of order 3 with no interior knots, but this is certainly not necessary. The mean and/or variance of the transition density could also depend on the time-fixed covariates  $A$  if desired.

In solving (3.13), we assume  $V_{1i}(t)$  and  $V_{2i}(t)$  are diagonal. This avoids estimating the variance matrix for the covariates when solving the estimating equation for the mean, and it also avoids the inefficient practice of estimating the fourth order moments of a multivariate normal. We simulate the underly-

ing longitudinal process on a discrete grid of time points and model the visit hazard rate as constant over each discretized interval; this discretization is not necessary because of the scope of  $\{H_i^{v,c}(t)\}$  in the conditional model (see Section 3.2.3), but it is helpful for comparisons that will be made with the next chapter. The choice of grid size does not seem to make much difference.

We take  $A_{1i} \sim \text{Uniform}(-20, 20)$ ,  $A_{i2} + 0.5 \sim \text{Bernoulli}(0.5)$ , and we use  $\theta_0^c$  consisting of:

$$\Sigma = \begin{pmatrix} 0.2 & 0.1 \\ 0.1 & 0.2 \end{pmatrix}.$$

In the OU model we use

$$\phi = \begin{pmatrix} 0.02 \\ 0.5 \end{pmatrix}, \quad \mu = \begin{pmatrix} 5 \\ 6 \end{pmatrix}, \quad \text{and} \quad B = \begin{pmatrix} 0.5 & 0 \\ 0 & 0.4 \end{pmatrix}.$$

and in the deterministic drift process we consider  $\mu(t, \beta) = \beta_0 - 0.5t + 0.03t^2$ . Note that  $\beta_0$  is not involved in (3.13) when  $\mu(t, \beta)$  is a B-spline with no interior knots, and so cannot (and does not need to be) estimated. In all tables, we report the average relative bias and the average asymptotic standard error (ASE) of each parameter estimate over 100 independent simulations, and we report the empirical standard deviation (ESE) of the 100 estimates. The estimate of the ASE for each simulation is calculated as the square root of  $n^{-1}$  times the diagonal of an estimate of (G.17) (see Appendix G.4).

Each simulation consisted of 100 subjects, each with an average of about 10 visits prior to the censoring time, which had a maximum of 35.

In Table 3.1, (3.13) with a correctly specified (3.9) is contrasted with a situation where the true diffusion variance is actually the heavy tailed multivariate  $t$  distribution with 3 degrees of freedom instead of the assumed normal dis-

Table 3.1: Simulation results for (3.13) with mean reverting drift using a correctly specified (3.9) and one with a true  $t_3$  diffusion variance

Parameter	(3.9)			var= $t_3$		
	rBias	ESE	ASE	rBias	ESE	ASE
$b_1$	0.008	0.038	0.038	0.010	0.045	0.045
$b_2$	0.006	0.031	0.031	0.023	0.037	0.035
$\mu_1$	0.000	0.023	0.023	0.000	0.023	0.024
$\mu_2$	0.000	0.029	0.027	-0.001	0.031	0.028
$\phi_1$	0.007	0.002	0.002	-0.001	0.002	0.002
$\phi_2$	-0.008	0.052	0.055	-0.009	0.050	0.057
$\Sigma_{11}^\dagger$	0.020	0.017	0.013	0.018	0.038	0.019
$\Sigma_{12}^\dagger$	0.012	0.010	0.009	-0.031	0.034	0.015
$\Sigma_{22}^\dagger$	0.012	0.017	0.016	0.025	0.052	0.028

tribution. Even with a misspecified variance, the relative biases remain quite low, with the biggest effect of misspecification seen in the variance parameter estimates and their standard errors.

Turning attention to the deterministic drift simulations, Table 3.2 again compares (3.13) with a correctly specified (3.9) to a situation where the true diffusion variance is actually the heavy tailed multivariate  $t$  distribution with 3 degrees of freedom. Here, there isn't even a significant increase in the bias and uncertainty of the variance parameter estimates with misspecification.

Table 3.2: Simulation results for (3.13) with deterministic drift using a correctly specified (3.9) and one with a true  $t_3$  diffusion variance

Parameter	(3.9)			var= $t_3$		
	rBias	ESE	ASE	rBias	ESE	ASE
$\beta_{11}$	-0.004	0.021	0.020	-0.002	0.030	0.037
$\beta_{12}$	0.001	0.001	0.001	0.001	0.001	0.002
$\beta_{21}$	0.003	0.017	0.020	-0.002	0.030	0.037
$\beta_{22}$	-0.003	0.001	0.001	0.001	0.001	0.002
$\Sigma_{11}$	0.019	0.022	0.012	-0.005	0.018	0.014
$\Sigma_{12}$	0.006	0.012	0.009	0.008	0.013	0.011
$\Sigma_{22}$	0.022	0.023	0.012	0.001	0.019	0.015

### 3.4 Data Analyses

#### 3.4.1 HUD-VASH

The marginal methods of Lin et al. (2004) and Bůžková and Lumley (2009) were used to analyze the HUD-VASH longitudinal dataset. For a thorough description of the data, including several exploratory plots, see Lin et al. (2004). The response of interest is the percentage of days a veteran spent homeless in the previous three months, and the covariate of interest is the intervention program, which has three levels: (i) full HUD-VASH intervention, which involved case management and housing vouchers (C+V); (ii) case management without vouchers (C); and (iii) standard care (std). Informative visitation is an issue here, and the visitation model is allowed to depend on intervention assignment (int), income at baseline (inc), indicator of social security or VA benefits at baseline

(SSVA), a Lehman measure of quality of life at baseline (QOL), and percentage of days homeless in the last three months (PH). Bůžková and Lumley (2009) use the previous value of PH because, as noted in Appendix E,  $dN(t)$  does not depend on  $X(t)$  in their model. Lin et al. (2004) and Bůžková and Lumley (2009) answer the question of which intervention program reduces homelessness most effectively by examining the marginal means; we do the same here. Additionally, our proposed conditional modeling procedure is used to study the relative volatilities of the intervention programs.

To these ends, we respectively model the marginal and conditional means of PH at time  $t$  as follows:

$$\mu(t) = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 I(\text{int} = C + V) + \beta_4 I(\text{int} = C), \quad (3.14)$$

$$\mu(t) = \text{PH}(s) + \beta_1(t - s) + \beta_2(t - s)\text{PH}(s) + \beta_3(t^2 - s^2) + \beta_4(t^2 - s^2)\text{PH}(s), \quad (3.15)$$

where  $s$  is the time of most recent visit prior to  $t$ . The standard intervention is therefore the reference level in the marginal mean and the effects of the full intervention and the case based intervention are compared to it, meaning  $\beta_3$  and  $\beta_4$  in (3.14) are of particular interest in the marginal setting. The degree two polynomial is sufficient to capture the overall trend in homelessness over the study period: it dips at first and then levels out. In the conditional setting, (3.15), we have allowed the rate of change of PH to depend on the most recently observed PH. The variance of PH in the marginal model is not the target of this analysis, so we haven't modeled it, but for the conditional model, the variance is modeled as:

$$\sigma^2(t) = \sigma_{\text{int}}^2 \log(1 + t - s),$$

where  $\sigma_{\text{int}}^2$  is different for each intervention, i.e.  $\text{int} = C + V, C$ , or  $\text{std}$ . By allowing each intervention to have its own variance, we can compare their relative

volatilities by comparing their variances. An alternative might have been to add terms to (3.15) that reflect an interaction between intervention and previous PH.

The marginal model requires estimation of the intensity of visitation using one of the following:

$$E(dN^*(t)) = \exp\left(\gamma^T(int, inc, SSVA, QOL, PH(s))\right) d\Lambda_0(t), \quad (3.16)$$

$$E(dN^*(t)) = \exp\left(\gamma^T(int, inc, SSVA, QOL, PH(t))\right) d\Lambda_0(t). \quad (3.17)$$

Equation (3.17) allows the visitation intensity at time  $t$  to depend on the response at time  $t$ , whereas (3.16) mimics Bůžková and Lumley (2009).

The resulting estimating equation is:

$$\sum_{i=1}^n \int_0^\tau \left( \frac{\frac{\partial \mu_i(t)}{\partial \beta^T} (PH_i(t) - \mu_i(t))}{e_{\text{int}_i} \left( (PH_i(t) - \mu_i(t))^2 - \sigma_i^2(t) \right)} \right) w_i^v(t, \hat{\gamma}) dN_i(t),$$

where  $e_{\text{int}_i}$  is a vector of zeros with a one in the appropriate row, and  $w_i^v(t, \hat{\gamma})$  is the inverse-intensity-of-visit-weight (which is equal to 1 in the conditional setting). The results for the marginal model are summarized in Table 3.3, and the results for the conditional model are summarized in Table 3.4. Table 3.3 estimates the ASE as the square root of  $n^{-1}$  times the diagonal of an estimate of (G.10) and Table 3.4 estimates the ASE as the square root of  $n^{-1}$  times the diagonal of an estimate of (G.17) (see Appendices G.2 and G.4).

All the marginal methods give the similar result that the full intervention is significantly better than either of the other two, but that the other two do not differ significantly. Differences between (3.16) and B-L are due to the different models for the reference level's mean PH over time.

As for the conditional model, using the correlations (not shown) it can be shown that  $SD(\hat{\sigma}_C^2 - \hat{\sigma}_{C+V}^2) \approx 23$ , so the full intervention is approximately two



Table 3.3: Marginal estimates of the intervention effects using the Lin-Ying method (no weighting), Buzkova-Lumley method, and the method proposed in this chapter with two types of weighting

	$\hat{\beta}_3$ (ASE)	$\hat{\beta}_4$ (ASE)
L-Y	-8.14 (3.00)	2.05 (4.98)
B-L	-10.56 (2.30)	0.32 (5.20)
(3.16)	-7.59 (2.28)	1.58 (3.50)
(3.17)	-7.65 (2.31)	1.69 (3.55)

Table 3.4: Conditional estimates of volatility of interventions

$\hat{\sigma}_{C+V}^2$ (ASE)	$\hat{\sigma}_C^2$ (ASE)	$\hat{\sigma}_{std}^2$ (ASE)
155 (13)	189 (19)	244 (19)

standard errors less volatile than the case based intervention, and obviously less volatile than the standard intervention. In other words, in addition to reducing the mean level of PH the most, the full intervention seems to be the best at preventing a homelessness relapse.

There are still questions that only the marginal models can answer, but the addition of the CEE to the toolbox adds an interesting and useful level of flexibility. It is of course also better for estimating individual subject trajectories, as seen in Figure 3.1.

### 3.4.2 Multivariate longitudinal protein

The HUD-VASH data of Section 3.4.1 illustrate the capability of the univariate marginal and conditional methodology, which is a special case of the method-

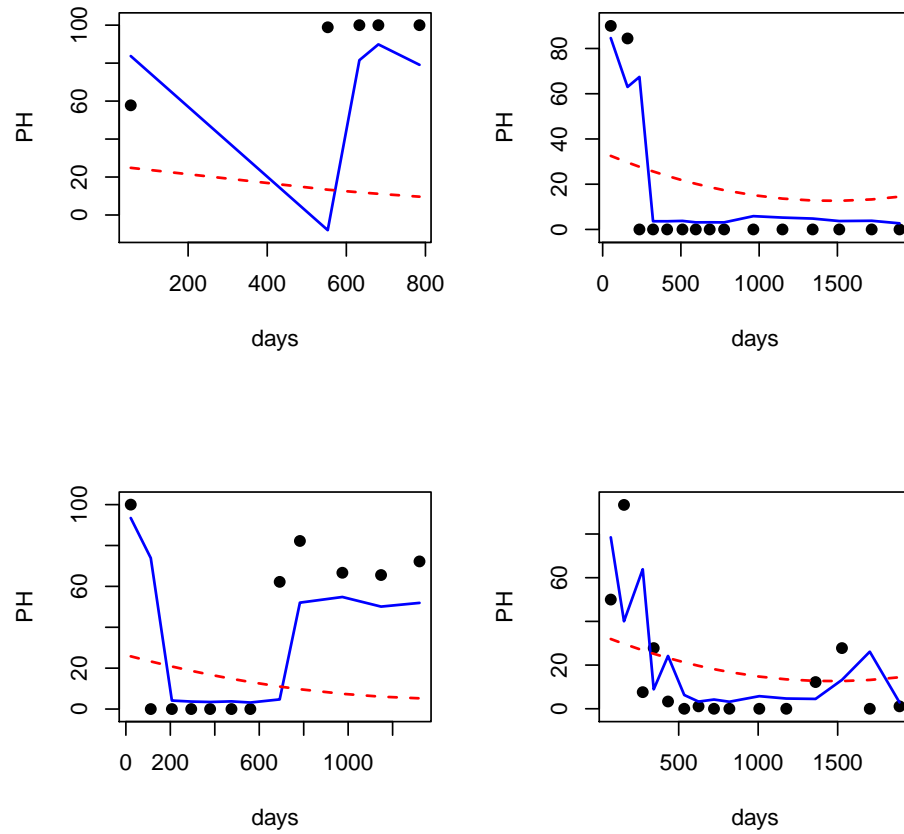


Figure 3.1: Percentage homelessness across time for four randomly selected subjects. The conditional estimates at each visit time are connected with the solid blue line; the marginal estimates are connected with the dashed red line

ology of Section 3.2. It is now desired to estimate correlation parameters in a multivariate data setting to fully display the range of our variance estimation methodology. To this end, we now introduce the longitudinal acute-phase protein measurement data, which were first presented in Kaysen et al. (2000). The paper by Kaysen and his coauthors focused more on the biological aspects of the data, so the more interesting comparison to make is with a later paper by Dubin and Müller (2005), which introduced the idea of dynamical correlation

to model multivariate longitudinal data in a computationally friendly manner, and used the longitudinal protein data to illustrate its methods.

Dynamical correlation relies on writing the longitudinal trajectory for each dimension of each subject as a linear combination of orthonormal basis functions. Correlations are then calculated for each subject and averaged over all subjects. As opposed to the usual definition of correlation, dynamical correlation calculates its measure of association by taking the expected cosine of the angle between standardized versions of these trajectories in Hilbert space. This cosine is based on functional inner products, and hence involves integration over time, meaning that, contrary to the term “dynamical”, the correlation between each pair of dimensions is not time-dependent. Dubin and Müller (2005) show that dynamical correlation has similar properties to the usual correlation, in particular its support is the interval from -1 to 1. In the case of a two-dimensional process, they note that “if both trajectories tend to be mostly on the same side of their time average (a constant), then dynamical correlation is positive; if the opposite occurs, then dynamical correlation is negative”.

Their method requires non-parametric smoothing to obtain complete trajectories when the data are only observed intermittently, but they note that correlation estimates are not sensitive to the choice of smoothing parameters.

The longitudinal protein data are indeed only observed intermittently, so Dubin and Müller (2005) require this smoothing step before correlation estimation in their data analysis. There are 34 subjects who have the complete multivariate set of five proteins measured a total of 611 times. The five proteins of interest are C-reactive protein (crp),  $\alpha$ -aminoglobulin (aag), ceruloplasmin (cer), transferrin (trf), and albumin (alb). The question of interest for Dubin and

Müller (2005), and for us, is what the signs of the pairwise correlations are: of particular biomedical interest to Dr. Kaysen and his colleagues is the hypothesis that *trf* and *alb*, which are known as negative acute-phase proteins (NAPPs), are negatively correlated with *crp*, *aag*, and *cer*, which are known as positive acute-phase proteins or simply acute-phase proteins (APPs).

Some basic multiple linear regressions showed that there is no effect of time on *crp* and *cer*, but that time is related linearly to *aag* and *trf*, and quadratically to *alb*. All these effects are quite small, and correlation estimates do not change significantly when these effects are not accounted for; nevertheless we propose the following model for conditional moments:

$$\begin{aligned}\mu(t) &= X(s) + (0, \beta_1(t-s), 0, \beta_2(t-s), \beta_3(t-s) + \beta_4(t^2 - s^2)), \\ \Sigma(t) &= \Sigma h(t-s),\end{aligned}\tag{3.18}$$

where  $\beta$  and  $\Sigma$  are respectively the 4-dimensional mean parameter and 15-dimensional variance parameter to be estimated,  $\mu$  is a column vector of conditional means of (*crp*, *aag*, *cer*, *trf*, *alb*) respectively, and  $h$  is a monotone function; we tried both identity and log, and did not get significantly different correlation estimates. We also tried letting rates of change of  $X$  depend on  $X(s)$  but this too did not change the results qualitatively, so we opted for the more parsimonious model. Unfortunately, the setup of (3.13) does not allow direct estimation of correlation because the empirical correlation estimate for any particular visit is always equal to one; this follows from the rank one form of the empirical variance estimate. Therefore the estimate of the variance matrix  $\Sigma$ , in (3.18), must be obtained from (3.13), and then used to calculate an estimate of the correlation matrix and its corresponding standard error.

Denoting the covariance between process  $j$  and process  $k$ , where  $j, k =$

crp, aag, cer, trf, alb, as  $\Sigma_{jk}$ , the corresponding correlation is

$$\rho_{jk} \equiv \frac{\Sigma_{jk}}{\sqrt{\Sigma_{jj}\Sigma_{kk}}}, \quad (3.19)$$

with an estimate obtained by replacing the quantities on the right side of (3.19) with their corresponding sample quantities obtained from (3.13). Then, defining  $\alpha_{jk} \equiv (\Sigma_{jj}, \Sigma_{jk}, \Sigma_{kk})$ , the asymptotic variance of the correlation estimate is obtained using the delta method:

$$\text{Var}(\hat{\rho}_{jk}) \equiv \frac{\partial \rho_{jk}}{\partial \alpha_{jk}} \text{Var}(\hat{\alpha}_{jk}) \frac{\partial \rho_{jk}}{\partial \alpha_{jk}^T}. \quad (3.20)$$

An estimate is obtained by replacing the quantities on the right side of (3.20) with their estimates: the derivative estimate is obtained from (3.13) and the variance estimate is  $n^{-1}$  times an estimate of the portion of (G.17) that relates to the variance parameters (see Appendix G.4).

The correlation estimates, based on (3.19), subscripted with their p-values for a two-sided test against zero (which assumes normality of the correlation estimates) are presented in Table 3.5. The dynamical correlation estimates, subscripted with bootstrapped p-values, for the Dubin and Müller (2005) model are shown in Table 3.6.

The signs of all the correlations in Table 3.5 agree with the hypothesis that NAPPs and APPs are negatively correlated over time, and almost all are highly significant; conversely, Table 3.6 has a positive correlation between cer and trf, which does not agree with the hypothesis, and it also has weaker significance elsewhere, particularly with the aag and trf correlation. Nevertheless, most of the correlation estimates in the two methods are qualitatively similar.

Table 3.5: Correlation parameter estimates based on (3.19) with their subscripted p-values for a two-sided test against zero

	crp	aag	cer	trf	alb
crp	1.000	0.585 <sub>(.000)</sub>	0.269 <sub>(.002)</sub>	−0.386 <sub>(.000)</sub>	−0.123 <sub>(.008)</sub>
aag	0.585 <sub>(.000)</sub>	1.000	0.669 <sub>(.000)</sub>	−0.381 <sub>(.000)</sub>	−0.125 <sub>(.019)</sub>
cer	0.269 <sub>(.002)</sub>	0.669 <sub>(.000)</sub>	1.000	−0.125 <sub>(.038)</sub>	−0.094 <sub>(.478)</sub>
trf	−0.386 <sub>(.000)</sub>	−0.381 <sub>(.000)</sub>	−0.125 <sub>(.038)</sub>	1.000	0.192 <sub>(.043)</sub>
alb	−0.123 <sub>(.008)</sub>	−0.125 <sub>(.019)</sub>	−0.094 <sub>(.478)</sub>	0.192 <sub>(.043)</sub>	1.000

Table 3.6: Dynamical correlation parameter estimates for the Dubin and Müller (2005) model with their subscripted bootstrapped p-values for a two-sided test against zero

	crp	aag	cer	trf	alb
crp	1.000	0.549 <sub>(.000)</sub>	0.387 <sub>(.024)</sub>	−0.215 <sub>(.072)</sub>	−0.298 <sub>(.004)</sub>
aag	0.549 <sub>(.000)</sub>	1.000	0.686 <sub>(.000)</sub>	−0.096 <sub>(.616)</sub>	−0.326 <sub>(.036)</sub>
cer	0.387 <sub>(.024)</sub>	0.686 <sub>(.000)</sub>	1.000	0.107 <sub>(.256)</sub>	−0.166 <sub>(.276)</sub>
trf	−0.215 <sub>(.072)</sub>	−0.096 <sub>(.616)</sub>	0.107 <sub>(.256)</sub>	1.000	0.247 <sub>(.060)</sub>
alb	−0.298 <sub>(.004)</sub>	−0.326 <sub>(.036)</sub>	−0.166 <sub>(.276)</sub>	0.247 <sub>(.060)</sub>	1.000

### 3.5 Discussion

This chapter has presented an innovative method for estimating parameters driving a multivariate longitudinal process with informative visitation. Several papers, most recently Bůžková and Lumley (2009), have used inverse-visitation-intensity-weighted marginal estimating equations to model a univariate mean; this chapter has generalized to a multivariate mean, and introduced a CEE that obviates the need to estimate weights. The CEE also relaxes the independent

censoring assumption, but it strengthens the visitation assumptions - only very slightly however. Simulations have shown that the CEE is robust to misspecification of the transition variance, and we have reanalyzed the popular HUD-VASH dataset from a conditional perspective, resulting in conclusions about the conditional volatility of the homelessness for the three different interventions. Specifically, we were able to conclude that the full intervention seems to do the best job of preventing a relapse into homelessness. We also reanalyzed the longitudinal protein data from Dubin and Müller (2005) and were able to show even more conclusively than they did that NAPPs are negatively correlated over time with APPs.

The conditional model from Section 3.2.3 has the advantage of providing the ability to estimate subject-specific trajectories that can be used in prediction (more detail on this will be provided in Chapter 5). In Chapter 4, missing response data will be considered, and the conditional specification will allow imputation of the missing dimensions of the response conditional on the observed dimensions, creating a novel estimating equation.

Two extensions to the current model would be the inclusion of a latent variable and the ability to deal with measurement error in the covariates. Similarly to Sun et al. (2007) and Liang et al. (2009), the inclusion of a latent variable could weaken the assumption regarding the relationship between the visitation and longitudinal processes. Independent and identically distributed measurement error with known variance, as in Tsiatis and Davidian (2001), could easily be included in (3.13). In models where  $\bar{X}_i(t^-)$  only enters the transition density mean linearly (e.g. the deterministic drift and the mean reverting drift), the expectation of the mean parameter estimates will not change, and in models where the

variance in the transition density does not depend on  $X$ , one can easily identify the variance due to the true variation and the variance due to measurement error.

Another interesting digression could be to model a discrete response using a continuous-time Markov chain (CTMC), while still allowing for informative visitation. The CTMC could be modeled conditionally or marginally depending on the setting.



## CHAPTER 4

# CONDITIONAL ESTIMATING EQUATIONS FOR MULTIVARIATE LONGITUDINAL DATA SUBJECT TO CENSORING, FAILURE, AND MISSINGNESS

### 4.1 Introduction

Chapter 3 considered a (weighted) estimating equation to estimate parameters indexing a censored longitudinal response process, allowing the (W)EE to be marginal or conditional. This chapter will focus on the conditional estimating equation (CEE) introduced in Section 3.2.3, but generalize it by introducing the possibilities of missingness in the response process and an end of study due to a terminal event (we will often just call it a “failure”, but the type of event can be more general).

Dealing with the introduction of failures is pretty straightforward: the conditional mean and variance specification in the CEE are now made conditional on survival as well as on the previous responses. This necessitates some new assumptions connecting the failure process with the censoring, longitudinal, and visitation processes.

The literature dealing with non-monotone missing responses (i.e. responses that can be missing at a particular time but observed at a later time) in multivariate longitudinal data seems limited. A recent paper by Aalen and Gunnes (2010), based on linear increments, addresses non-monotone multivariate missingness but assumes the missingness for a subject at a particular time is either all or nothing, and only considers observations of a continuous process in discrete

time without informative visitation. Two popular solutions, which might handle the more general setting we consider, are last value carried forward (LVCF) and multiple imputation. But LVCF is relatively naïve, and like using mean values to replace the missing values, it obviously leads to bias of the regression parameter estimates. Multiple imputation (a good review is available in Rubin, 1996), might be feasible but the computational time required for such a Monte Carlo approach is a detriment, and it turns out a small adjustment of our CEE from Chapter 3 provides a clean, analytic solution to parameter estimation in the presence of missingness.

The two main assumptions, which were not made in Chapter 3 but are made now, are that the missing data are missing at random (MAR, Rubin, 1976), actually more precisely sequentially missing at random (S-MAR, see Hogan et al., 2004, Robins et al., 1995), and that the longitudinal process follows a normal transition density. The former is usually quite reasonable, and says that missingness depends only on previously and currently observed data and not on the missing values themselves. Whenever the decision to measure a particular dimension of the response is made by a doctor or nurse, for example, it is being made based only on observed values. This S-MAR assumption combined with the correct specification of a parametric transition density of the longitudinal process results in consistent estimates of the regression parameters because as pointed out in Lipsitz et al. (1999), consistent estimates follow from either correct specification of the missing data mechanism or of the distribution of the missing data given the observed data, the latter of which is satisfied in this chapter. The choice of normality for this parametric assumption makes this projection of the unobserved data onto the observed data computationally convenient. As Tsiatis et al. (1995) explain, this normality assumption is “not technically rea-

sonable as it would necessitate the existence of [longitudinal] and failure time processes that induce the family of joint Gaussian distributions conditional on being at risk at each time  $t$ ” but it is “practically reasonable” because a normality assumption is usually quite robust.

This chapter will proceed as follows. Section 4.2.1 deals with a longitudinal process subject to a terminal event, Section 4.2.2 deals with a longitudinal process subject to missingness, and Section 4.2.3 provides the resulting CEE to obtain consistent and asymptotically normal estimates of the transition density parameters. Section 4.3 shows some simulation results, which expand on those from Section 3.3 in the previous chapter. Section 4.4 provides an analysis of medical cost data to illustrate modeling a longitudinal process of interest when it is subject to a terminal event. Finally, Section 4.5 wraps up the chapter and presents some possible future research directions. Technical details and large sample theory are provided in Appendices H-J.

## 4.2 Methodology

### 4.2.1 Introducing failure times

With the inclusion of failure times,  $T_i$ , the following changes are made from the notation in Section 3.2.1. The at-risk indicator is now  $\xi_i(t) = I(C_i \geq t)I(T_i \geq t)$ , with subject  $i$ ’s visits coming at  $0 \leq t_{i0} < t_{i1} < \dots < t_{iK_i} < \Upsilon_i < \tau$ , where  $\Upsilon_i = \min(C_i, T_i)$  is observed along with  $\delta_i = I(T_i \leq C_i)$ . Also,  $N_i(t) = N_i^*(\min(t, \Upsilon_i))$

counts the number of observed visits, and (3.8) and (3.9) are replaced by

$$E\left(X_i(t)|H_i^{m,c}(t), T_i \geq t\right) = \mu\left(h_i^{m,c}(t), \theta_0\right), \quad (4.1)$$

$$Var\left(X_i(t)|H_i^{m,c}(t), T_i \geq t\right) = \Sigma\left(h_i^{m,c}(t), \theta_0\right). \quad (4.2)$$

Tsiatis et al. (1995) use (4.1) and (4.2), including a normality assumption, in a two stage approach to survival parameter estimation, and Fine et al. (2004), while considering time-indexed parameters in a functional generalized linear model to marginally describe a longitudinal process in presence of a terminal event, use (4.1) without specifying a distribution, a much weaker assumption. When the probability of death is relatively low, the difference between (4.1)-(4.2) and (3.8)-(3.9) is quite small (see Sections 4.3 and 4.4), and these new assumptions avoid the cumbersome calculation of inverse-probability-of-survival-weights which would be required for an unbiased CEE when assumptions (3.8)-(3.9) are made. Assumptions (4.1)-(4.2) also facilitate the calculation of the distribution of the longitudinal process at an arbitrary point in time, conditioned on its past values, a calculation we will make extensive use of in Chapter 5.

## 4.2.2 Introducing missingness

The possibility that only a subset of the process is observed at each visit time complicates the parameter estimation in two ways. First, the dimensions of the process might not all have been observed at the most recent visit, meaning that the mean and variance of the transition density assumed in Section 3.2.3 will not be valid without some adjustments. And second, the dimensions might not all be measured at the current visit, meaning that, for maximal efficiency, the conditional distribution of the missing data given the observed data must be solved

for. By assuming that the measurements are missing at random (MAR, see Rubin, 1976), actually more precisely sequentially missing at random (S-MAR, see Hogan et al., 2004, Robins et al., 1995), i.e. the missingness mechanism depends only on previously and currently observed data, the conditional distribution of the missing data given the observed data follows from a parametric assumption on the transition density of the longitudinal process. As pointed out in Lipsitz et al. (1999), consistent parameter estimates follow from either correct specification of the missing data mechanism or of the distribution of the missing data given the observed data, the latter of which is satisfied here (note that this departs from the usual limitations of MAR data in GEE, see Zorn (2001) for example, because a parametric distribution has now been assumed). This conditional distribution calculation is made available in closed form by assuming that the transition density of the longitudinal process follows a multivariate normal distribution, which will be done in everything that follows. That is, we assume that  $(X_i(t)|H_i^{m,c}(t), T_i \geq t)$  follows a multivariate normal distribution, the mean and variance of which are specified in (4.3) and (4.4) respectively. We now address the two aforementioned complications in more detail.

Calculating the distribution of  $X_i(t)$  conditioned on  $\{H_i^{m,c}(t), T_i \geq t\}$ , where some of the previous response measurements may now be missing, is a non-trivial task which is based on (4.1) and (4.2). The result is:

$$E(X_i(t)|H_i^{m,c}(t), T_i \geq t) = \tilde{\mu}(h_i^{m,c}(t), \theta_0) \equiv \tilde{\mu}_i(t), \quad (4.3)$$

$$Var(X_i(t)|H_i^{m,c}(t), T_i \geq t) = \tilde{\Sigma}(h_i^{m,c}(t), \theta_0) \equiv \tilde{\Sigma}_i(t). \quad (4.4)$$

The derivations of  $\tilde{\mu}_i(t)$  and  $\tilde{\Sigma}_i(t)$  will of course depend on the original form of the functions  $\mu$  and  $\Sigma$ . They also depend on  $\theta$ , but to avoid notational clutter, we've suppressed that dependence. The derivations for a multivariate Ornstein-

Uhlenbeck process are shown as an example in Appendix H.

To obtain the missing data/failure time extension of (3.13), the remaining requirement is to find the conditional distribution of  $X_i^u(t)$  given  $\{X_i^o(t), H_i^{m,c}(t)\}$ , where  $X_i^u(t)$  and  $X_i^o(t)$  are the dimensions of the longitudinal process that were unobserved and observed respectively at time  $t$  (with the minor modification that to be in  $X_i^o(t)$ , the dimension must have been observed once previously). Note that  $T_i \geq t$  is implicit with the observation of  $X_i^o(t)$ . Define the partitions

$$\tilde{\mu}_i(t) = \begin{pmatrix} \tilde{\mu}_{i1}(t) \\ \tilde{\mu}_{i2}(t) \end{pmatrix} \quad \text{and} \quad \tilde{\Sigma}_i(t) = \begin{pmatrix} \tilde{\Sigma}_{i11}(t) & \tilde{\Sigma}_{i12}(t) \\ \tilde{\Sigma}_{i21}(t) & \tilde{\Sigma}_{i22}(t) \end{pmatrix},$$

where the subscript 1 corresponds to the unobserved portion of  $X_i(t)$  and the subscript 2 corresponds to the observed portion. The required conditional distribution is easily found now because by the well known property of the multivariate normal,

$$\begin{aligned} & \left( X_i^u(t) | X_i^o(t) = c, H_i^{m,c}(t) \right) \\ & \sim N \left( \tilde{\mu}_{i1}(t) + \tilde{\Sigma}_{i12}(t) \tilde{\Sigma}_{i22}^{-1}(t) (c - \tilde{\mu}_{i2}(t)), \tilde{\Sigma}_{i11}(t) - \tilde{\Sigma}_{i12}(t) \tilde{\Sigma}_{i22}^{-1}(t) \tilde{\Sigma}_{i21}(t) \right). \end{aligned} \quad (4.5)$$

### 4.2.3 CEE with failure times and missingness

In order to consistently estimate  $\theta_0$ , a coarsening-at-random-type (Heitjan and Rubin, 1991; Gill et al., 1996; Gill and Grünwald, 2008) assumption regarding the lifetime process and the associated censoring process is required. Instead of (3.10), this chapter requires

$$E \left( X_i(t)^{\otimes j} | H_i^{m,c}(t), Y_i \geq t \right) = E \left( X_i(t)^{\otimes j} | H_i^{m,c}(t), T_i \geq t \right) \quad (4.6)$$

for all  $t$ , where  $j = 1, 2$  and  $x^{\otimes 1} = x$  and  $x^{\otimes 2} = xx^T$  for a vector  $x$ . In our proposed conditional model this is weaker than the “sequential ignorability of censoring”

assumption defined in Scharfstein and Robins (2002) in that the hazard for the censoring process is allowed to depend on  $T_i \geq t$  even after conditioning on the history of covariates and responses. We also need the analogs of (3.11) and (3.12), which are

$$E\left(dN_i^*(t)|H_i^{m,c}(t), H_i^{v,c}(t), X_i(t), \Upsilon_i \geq t\right) = E\left(dN_i^*(t)|H_i^{v,c}(t)\right), \quad (4.7)$$

$$E\left(X_i(t)^{\otimes j}|H_i^{m,c}(t), H_i^{v,c}(t), \Upsilon_i \geq t\right) = E\left(X_i(t)^{\otimes j}|H_i^{m,c}(t), \Upsilon_i \geq t\right), \quad (4.8)$$

where  $j = 1, 2$ . It is important to emphasize here that  $\{H_i^{v,c}(t)\}$  only contains observed information. Hence, the visitation intensity cannot depend on missing data.

The following CEE handles parameter estimation for longitudinal data subject to failures and missingness:

$$\tilde{U}^c(\theta) = \sum_{i=1}^n \int_0^\tau \left( \frac{\partial \tilde{\mu}_i(t)}{\partial \beta^T} \tilde{V}_{1i}^{-1}(t) \left( E(X_i(t)|X_i^\dagger(t)) - \tilde{\mu}_i(t) \right) \right) f(h_i^{m,c}(t)) dN_i(t), \quad (4.9)$$

where  $X_i^\dagger(t) \equiv \{X_i^o(t), H_i^{m,c}(t)\}$ , and  $\tilde{s}_i^*(t)$  is the upper triangle of  $(X_i(t) - \tilde{\mu}_i(t))(X_i(t) - \tilde{\mu}_i(t))^T$ . Calculations of the expected value terms in (4.9) are easily carried out using (4.5), and are provided in Appendix I. Note that if  $X_i^\dagger(t)$  does not contain any information about  $X(t)$  then  $E(X_i(t)|X_i^\dagger(t)) = \tilde{\mu}_i(t)$  and  $E(\tilde{s}_i^*(t)|X_i^\dagger(t)) = \tilde{\Sigma}_i^*(t)$ , and so the contribution of that term to (4.9) would be zero, regardless of the value of  $dN_i(t)$ .

**Theorem 4.2.1.** Given the normality of  $X$  defined by (4.3)-(4.4), and given (4.6)-(4.8) and regularity conditions (J1)-(J5) in Appendix J.1, the  $\hat{\theta}$  that solves  $0 = \tilde{U}^c(\hat{\theta})$  is a consistent and asymptotically normal estimator of  $\theta_0$ , the true parameter.

The proof of Theorem 4.2.1 is provided in Appendix J.2, along with an estimate of the asymptotic variance of  $\hat{\theta}$ .

#### 4.2.4 IPSW digression

Even assuming no missingness, assumptions (4.1) and (4.2), which were introduced in Section 4.2.1 were vital to the form of (4.9). As mentioned in Section 4.2.1, if (3.8)-(3.9) were to be used instead, an inverse-probability-of-survival-weight (IPSW) would be required in the CEE to create unbiasedness.

Focusing on the mean estimating equation, with the variance portion following the exact same logic, it is easy to show using the details in Appendix J.2 that unbiasedness without (4.1) and (4.2) would require

$$E(w(t)I(T \geq t)(X(t) - \tilde{\mu}(t))|H^{m,c}(t)) = 0 \quad (4.10)$$

for all subjects and all  $t$  for some weight function  $w(t)$ . Ignoring for the moment that  $X(t)$  is an internal covariate, i.e. its existence at time  $t$  implies  $T \geq t$ , the requirement (4.10) can be changed to

$$E((X(t) - \tilde{\mu}(t))w(t)P(T \geq t|H^{m,c}(t), X(t))|H^{m,c}(t)) = 0 \quad (4.11)$$

if  $w(t)$  depends only on  $\{H^{m,c}(t), X(t)\}$ . One such  $w(t)$  would be

$$\frac{1}{P(T \geq t|H^{m,c}(t), X(t))}. \quad (4.12)$$

Given that  $X(t)$  is actually an internal covariate, this suggests that rather than exactly following this form,  $w(t)$  is the inverse of the probability of a future subject following this same trajectory surviving all the way from the previous visit time to time  $t$ . This is in fact exactly the logic behind inverse-probability-weighting: the contribution of the subject that did survive to time  $t$  needs to be up-weighted to account for all the other similar subjects that did not survive (Schoop, 2008).

Actually calculating this weight is not easy. It requires a strong assumption on the joint distribution of the longitudinal and survival processes and it will



be computationally intensive because survival parameters must be estimated simultaneously with the longitudinal parameters of interest. And even then, some numerical integration over a grid of  $X$  values up to time  $t$  is required just to approximate the weight. It is for these reasons that we opt for assumptions (4.1) and (4.2) in general. Section 4.4 will discuss a situation where IPSW is feasible.

### 4.3 Simulations

This chapter has introduced both failures and missing responses. It is helpful to introduce them one at a time to check for a possible introduction of bias, so failures will be introduced first without any missingness.

See Section 3.3 for a description of the setup for the response, visitation, and censoring. The differences between Section 3.3 and this section are that failures are now simulated to create approximately 50% censoring and approximately 15% fewer observations per subject, and that the ASE estimate is now calculated as the square root of  $n^{-1}$  times the diagonal of an estimate of (J.7) (see Appendix J.2).

Although (4.1) and (4.2) are used as model assumptions, there is no obvious way to simulate the data conditioned on at-risk status, so the data in this section are simulated to have a marginal normal distribution; that is, as specified by (3.8) and (3.9), meaning that we are studying the performance of (4.9) under this misspecification in all the simulations in this section.

For the final two columns of both Table 4.1 and Table 4.2, about 30% of the re-

Table 4.1: Simulation results for mean reverting drift using (4.9). The three columns respectively represent no missingness, 30% missingness, and 30% missingness with a misspecified variance

Parameter	0%, var=normal			30%, var=normal			30%, var= $t_3$		
	rBias	ESE	ASE	rBias	ESE	ASE	rBias	ESE	ASE
$b_1$	0.024	0.041	0.042	0.035	0.056	0.052	0.021	0.050	0.050
$b_2$	0.019	0.033	0.035	0.020	0.044	0.041	0.024	0.045	0.041
$\mu_1$	-0.001	0.022	0.027	0.000	0.034	0.030	-0.001	0.032	0.029
$\mu_2$	0.000	0.034	0.033	0.000	0.036	0.035	-0.002	0.034	0.036
$\phi_1$	0.006	0.003	0.002	0.009	0.003	0.003	0.001	0.002	0.003
$\phi_2$	0.006	0.062	0.066	-0.006	0.081	0.071	-0.011	0.075	0.072
$\Sigma_{11}^\dagger$	0.000	0.013	0.013	-0.013	0.015	0.015	-0.016	0.019	0.017
$\Sigma_{12}^\dagger$	-0.010	0.010	0.010	-0.029	0.030	0.022	-0.035	0.025	0.023
$\Sigma_{22}^\dagger$	0.002	0.017	0.018	0.000	0.022	0.020	-0.008	0.038	0.025

sponses are deleted using a MAR approach, with missingness more likely when previously observed responses were high. In the final column of both Table 4.1 and Table 4.2, the true diffusion variance is simulated as the heavy tailed multivariate  $t$  distribution with 3 degrees of freedom instead of the assumed normal distribution.

Table 4.1 presents the results for the mean reverting drift model. All the relative biases are quite low, even for the misspecified variance model. The misspecification resulting from (4.1) and (4.2) seems to cause a slight positive bias in  $\hat{b}$ , but considering that failures are simulated with a relatively high rate, the bias is not large (in simulations not shown, we saw that the bias does not seem to change significantly when the visitation rate is changed). The other noticeable thing is that the variance parameter estimation gets worse, particularly the

Table 4.2: Simulation results for deterministic drift using (4.9). The three columns respectively represent no missingness, 30% missingness, and 30% missingness with a misspecified variance

Parameter	0%, var=normal			30%, var=normal			30%, var= $t_3$		
	rBias	ESE	ASE	rBias	ESE	ASE	rBias	ESE	ASE
$\beta_{11}$	-0.004	0.025	0.035	-0.010	0.027	0.029	0.006	0.026	0.030
$\beta_{12}$	0.001	0.001	0.002	-0.004	0.001	0.001	-0.009	0.001	0.001
$\beta_{21}$	-0.005	0.025	0.035	-0.003	0.026	0.030	0.000	0.031	0.029
$\beta_{22}$	-0.002	0.001	0.002	-0.002	0.001	0.001	-0.007	0.001	0.001
$\Sigma_{11}$	0.010	0.012	0.011	0.002	0.013	0.013	-0.014	0.018	0.015
$\Sigma_{12}$	0.005	0.010	0.008	0.024	0.043	0.030	-0.035	0.033	0.029
$\Sigma_{22}$	0.004	0.012	0.011	0.012	0.015	0.013	-0.008	0.021	0.017

$\Sigma_{12}^\dagger$  estimate, in terms of bias and efficiency, when missingness is introduced and then slightly worse again when misspecification is introduced. This is not surprising because with missingness there is far less information available to estimate correlation between the two dimensions. With larger sample sizes the bias in  $\hat{\Sigma}_{12}^\dagger$  is greatly decreased for the column corresponding to missingness with correctly specified variance.

Table 4.2 presents the results for the deterministic drift model, and the relative biases are again quite low in all cases. The same increase in bias and variance of the  $\Sigma$  estimates, particularly  $\hat{\Sigma}_{12}$ , is observed across the table. And as expected, the  $\beta$  estimates have a small negative bias due to the use of (4.1) and (4.2); this bias is created because  $E(X(t)|H^{m,c}(t), T \geq t) < E(X(t)|H^{m,c}(t))$  when low  $X$  values lead to fewer failures, as was the case for these simulations.

## 4.4 Data Analysis: Medical costs

In most applications involving longitudinal and time-to-event data, the longitudinal process is of limited interest on its own, and is only modeled as a means to survival modeling. However, with medical cost data, while survival is still obviously more important, there is also interest in modeling the longitudinal patterns of medical costs incurred by patients. This section will consider such a dataset, and will model the medical cost in the current month as a function of previous monthly costs and other covariates.

### 4.4.1 CHF data introduction

This dataset involves longitudinal monthly medical costs of 1397 chronic heart failure (CHF) patients taken from the clinical data repository at the University of Virginia (UVa), and was previously analyzed in Liu et al. (2008a) and in Liu et al. (2008b). The reader is referred to those papers and the references therein for a more detailed review of the medical cost modeling literature, as well as for a more complete description of the CHF dataset. We now give the important details of the dataset.

This class of dataset is quite different than the one in, for example, Section 3.4.2, because the response of interest there was a continuous underlying process, which was inevitably unknown between the visit times; conversely, this process is tabulated only monthly, and daily data were apparently not of any interest. Therefore, although there is informative visitation (costs are higher for subjects who visit than for those who do not visit), the underlying longitudinal

process is known to be zero for months where no visits take place. Recall from Section 4.2.4 that the roadblock to the use of the cleaner assumption (3.8) instead of (4.1) was the complete inability to estimate an IPSW without jointly modeling the failure and response processes. But since only the monthly response data are of interest in this setting, calculation of the IPSW is more straightforward (because no interpolation of a continuous process is required), and the estimates obtained using (3.13) with assumption (3.8) and weight based on (4.12) can be contrasted with the estimates obtained using (3.13) with assumption (4.1) and no IPSW. Recall that Section 4.3 showed that the conditional mean parameter estimates for the two methods are quite similar.

For 41% of the months studied (the data for the first month of study were removed from the original dataset because virtually all patients had costs in their first month, see Liu et al., 2008a), there was no monetary cost, i.e. the patient did not return for follow-up. The months with non-zero cost are classified as inpatient or outpatient months, with inpatient months accumulating a much larger cost. The costs are skewed to the right, so log transforms of  $\text{cost}+1$  and of  $\text{cost}$  were respectively taken in this section and by Liu and his colleagues. In addition to the monthly monetary cost to the UVa health system and the inpatient indicator, the other available data are the time-fixed covariates of age (centered at 72), gender (indicator of male), and race (indicator of white). The mean follow-up time is 18.5 months ( $\text{SD} = 9.0$ ), and 16% of patients died before the end of study, while the remaining 84% were censored.

#### **4.4.2 Previous models considered by Liu and coauthors for the CHF data**

Liu et al. (2008a) use four generalized linear models (GLMs) to model monthly indicator of visit, a monthly indicator of inpatient conditional on visit equal one, a monthly inpatient cost conditional on visit equal one and inpatient equal one, and a monthly outpatient cost conditional on visit equal one and inpatient equal zero. The four GLMs are connected by correlated random effects to account for between-subject heterogeneity. Liu et al. (2008b) broaden the analysis by modeling three processes - visitation, cost, and lifetime - using a joint random effects model. As in Liu et al. (2008a), they model the monthly cost conditional on visiting in that month. One implication of the joint random effects models in these papers is that regression coefficients are considered subject-specific.

#### **4.4.3 A new model for the CHF data**

In applying our CEEs (both with and without IPSW) we chose to model the cost of the month without conditioning on visit status. This creates a bimodal residual distribution (because of the zero costs), but has the more direct interpretation of actual cost for the month. Differing from the papers of Liu and his coauthors, which incorporate all between-subject heterogeneity (i.e. frailty) in their cost model through either one or two random effects, we condition on the average of previous costs, which incorporates the between-subject heterogeneity by using covariate information that changes over the course of observation (Aalen et al., 2004). We also condition on the inpatient indicator (inpat) for the *current* month. This conditioning is possible because as mentioned in Section

3.2.3,  $A(k)$  (inpat(k) in this case) can be used to model  $X(k)$  (the cost for the  $k^{\text{th}}$  month in this case). Due to the fact that costs associated with inpatients are where the vast majority of total expenditures lie, separate modeling of inpatient months and non-inpatient months is highly desirable.

Importantly, the cost for the final (incomplete) month is actually available for the CHF data, and if it is taken into account, an average cost per day analysis using assumption (3.8) without any IPSW could be undertaken. This contrasts the CHF dataset with most applications that are subject to a terminal event because usually the longitudinal process is not observed in the interval immediately preceding death, creating the need for an IPSW (or for modeling assumption (4.1)). We however want to contrast the use of IPSW to the case with no IPSW, and so we ignore this final month cost in our models for both mortality and cost. Another reason to exclude this final month is that we do not have information regarding the pattern of cost accumulation within months; e.g. in a censored month with an outpatient visit, the expected cost for the censored part of the month might be very close to zero.

We use assumption (4.6), and for the mortality, assume an intensity driven by a Cox model:

$$E\left(dN^{d*}(t)|H^d(t)\right) = \exp\left(\kappa_0^T(\overline{\text{cost}}_3(t), \text{age}, \text{male}, \text{white}, \text{inpat}(t))\right)d\Lambda_0^d(t), \quad (4.13)$$

where  $N^{d*}(t)$  is the counting process for death and  $H^d(t)$  is the filtration it depends on (see Chapter 5 for a much more detailed discussion of survival modeling), inpat(t) is the inpatient indicator for the most recent complete month, and  $\overline{\text{cost}}_j(t)$  denotes the average cost over the previous  $j$  complete months. We considered  $j = 1, \dots, 10$ , but  $j = 3$  had the highest Cox partial likelihood. We do not take into account visitation history, as suggested in Liu et al. (2008b), because

its effect is insignificant when cost history is accounted for. For the conditional mean cost of the  $k^{th}$  month we assume:

$$\mu(k) = \beta_0^T (1, k, k^2, \overline{\text{cost}}_3(k), \text{age}, \text{age}^2, \text{inpat}(k)). \quad (4.14)$$

The model in (4.14) allows evaluation of the effects of the time-fixed covariates conditional on previous cost and inpatient status. It is similar to the model in Liu et al. (2008b) in that it models quadratic effects of age and time: the theory regarding quadratic age is that the oldest patients are treated less aggressively, and hence more cheaply, and the theory regarding quadratic time is that there is a high initial cost due to diagnosis and early treatment, and then costs decrease and flatten out before increasing shortly before death. This is known as the “bathtub” effect. We do not however use male as a predictor because when conditioning on inpat, there is no significant gender difference; the same result is found in Liu et al. (2008a). We also do not use white as a predictor because after conditioning on  $\overline{\text{cost}}_3(k)$ , it is insignificant: controlling for the differing cost trajectories of the different races wipes out their predictive value for the current cost.

We consider two realizations of (4.14), one with  $\mu(k)$  defined as (3.8) and one with  $\mu(k)$  defined as (4.1). This implies there are actually two different  $\beta_0$  vectors being estimated by the two CEEs; we denote the former as  $\beta_0^{(1)}$  and the latter as  $\beta_0^{(2)}$ .

Before considering the CEEs for estimating the conditional mean monthly cost, (4.13) is fit using a Cox model to estimate  $\kappa_0$ . The results are shown in Table 4.3. There are significant effects for  $\overline{\text{cost}}_3(k)$ , age, and inpat(k), but the effects of male and white are not quite significant at the 0.05 level.

The parameter estimates obtained using (4.14) in (3.13) are presented in Ta-



Table 4.3: Estimate of  $\kappa_0$  and its ASE in (4.13) for the CHF data

covariate	$\hat{\kappa}_0$ (ASE)
$\overline{\text{cost}}_3$	0.1322 (0.0241)
age	0.0424 (0.0078)
male	0.1953 (0.1217)
white	-0.2103 (0.1325)
inpat	1.0777 (0.1766)

ble 4.4, where it is evident that the use of assumption (3.8) with IPSW instead of assumption (4.1) with no IPSW, does not make a significant difference in estimates of  $\beta_0$ . By conditioning on previous cost instead of using a subject-specific random effect (as in Liu et al., 2008b), the subject-specific effect on cost is effectively allowed to change over time; despite this different model, the conclusions reached here are quite similar to those of Liu et al. (2008b): there is evidence of a “bathtub” cost shape over time, and there is some evidence of quadratic effect of age. As expected, inpatient indicator has quite a large effect on cost. But as mentioned earlier, male and white are insignificant when included in (4.14) with inpatient status and previous costs already accounted for.

For an analysis involving data subject to missingness as well as informative visitation, censoring, and the presence of a terminal event, the reader is referred to the analysis of cardiac data in Section 5.5.

Table 4.4: Estimates of  $\beta_0^{(1)}$  and  $\beta_0^{(2)}$ , and their ASEs, from (4.14) for the CHF data. The second column corresponds to the use of assumption (3.8) with IPSW in (3.13); the final column corresponds to the use of assumption (4.1) without any IPSW in (3.13)

covariate	$\hat{\beta}_0^{(1)}$ (ASE)	$\hat{\beta}_0^{(2)}$ (ASE)
intercept	1.5020 (0.0783)	1.4458 (0.0762)
month	-0.0294 (0.0097)	-0.0249 (0.0094)
month <sup>2</sup>	0.0012 (0.0003)	0.0011 (0.0003)
$\overline{\text{cost}}_3$	0.4919 (0.0142)	0.4970 (0.0143)
age	0.0064 (0.0035)	0.0061 (0.0034)
age <sup>2</sup>	-0.0004 (0.0004)	-0.0004 (0.0004)
inpat	3.8202 (0.0932)	3.8192 (0.0907)

## 4.5 Discussion

This chapter helps to fill a hole in the literature: modeling continuous multivariate longitudinal processes that are subject to missingness. Specifically, it has generalized Chapter 3 by estimating parameters driving a multivariate longitudinal process of interest that is not only subject to censoring, but also to missingness and a terminal event. Handling the failures is relatively easy due to some assumptions connecting the failure time process to the longitudinal, visitation, and censoring processes. Handling the missingness is more involved and consistency of parameter estimates requires a sequential MAR assumption and a parametric assumption on the response process to derive the distribution of the missing data conditioned on the observed data (it is for this reason that the methods described in this chapter do not apply to marginal models - once a parametric transition density is assumed, the estimating equation becomes

conditional).

Simulations have shown that the proposed model for missing data performs very well in practice, and that the CEE is robust to misspecification of the transition variance. They have also shown that despite modeling normality conditional on at-risk status, parameters corresponding to an unconditional normal are still estimated with very small bias.

We have reanalyzed the medical cost data from Liu et al. (2008a) and Liu et al. (2008b), handling between-subject heterogeneity by conditioning on past events rather than using random effects. Supporting the conclusions from Liu et al. (2008b), our model showed quadratic effects of age and time on the medical costs, and supporting the conclusion from Liu et al. (2008a) we found that there are no gender differences when inpatient status has been accounted for. Our model also showed that conditional on past costs, whites do not have lower current costs: controlling for the differing cost trajectories of the different races wipes out their predictive value for the current cost.

We have assumed that some transformation of the longitudinal process follows a multivariate normal transition density. But future work might consider copula methods instead, which would enable direct modeling of a much wider family of distributions. The computational simplicity of our method would be mitigated though. Nevertheless, this could be an area that deserves further exploration.

Although the methods of this chapter are not appropriate for marginally modeling an intermittently observed response subject to missingness, they could be utilized to impute a covariate process subject to missingness, which

could then be used in a marginal estimating equation for a response without missingness.

## CHAPTER 5

### SURVIVAL PREDICTION BASED ON DISCRETELY OBSERVED COVARIATES WITH MISSINGNESS

#### 5.1 Introduction

Chapter 4 considered a conditional estimating equation (CEE) to estimate parameters indexing a censored multivariate longitudinal process that was subject to missingness and an end of study due to a terminal event. This chapter will utilize that methodology by employing a variation of a Cox model to predict survival that depends on the underlying longitudinal process. In Chapters 3 and 4, the longitudinal process was of central interest, but now the survival prediction is of central interest, so hereafter we will refer to the longitudinal process as the “covariate process”. This chapter was inspired by data from a cardiac care unit in Ann Arbor, Michigan, provided to us by Dr. Mark Cowen. These data have several biological processes that are measured intermittently with missingness, and the ultimate goal is to effectively assign hospital resources to the patients in the greatest danger of immediate death.

The assumption made in Chapter 4 regarding the Gaussian transition density of the longitudinal process, which helped handle missingness, will also allow calculation of the distribution of the covariate trajectories at arbitrary times, while making use of the complete subject histories. Once this distribution is known, it allows the utilization of a modification of a Cox partial likelihood approach to estimate regression parameters in the failure hazard. If assumptions (4.1) and (4.2) hold, consistency of the Cox model estimates can be shown (this is done in Appendix K.2), but as discussed in Tsiatis et al. (1995), these as-

sumptions, while completely practical, may not be theoretically feasible: they “necessitate the existence of covariate and failure time processes that induce the family of joint Gaussian distributions conditional on being at risk at each time  $t$ .” Hence, we compare this partial likelihood method to a similar method proposed by Tsiatis et al. (1995). These two approaches will be described in detail in Section 5.3. Without missing data, it would be feasible to choose another parametric distribution for the covariates as long as its moment generating function is known: similarly to the Cox model, covariates enter our model only through the identity and exponential functions. Self and Pawitan (1992) actually propose a relative risk function that is linear in the covariates, so only the expectation of the covariates would be required in this case.

Prediction of survival to time  $s+t$  conditioned only on information up to time  $s$ , without parametric assumptions on covariates, has been studied by Zheng and Heagerty (2005) and Van Houwelingen (2007) with the former terming it “partly conditional modeling”, and the latter calling it “landmark analysis”. However, having already made a parametric assumption about the covariates in Section 4.2.2, there is enough information about the covariate process that conditioning up to time  $s$  can still elucidate  $X(u)$  for  $u > s$  to some degree.

The approach to the prediction problem that is presented in this chapter differs from most of the literature because we do not jointly model the longitudinal and time-to-event processes by connecting them with an unobserved latent variable. An excellent survey of joint modeling is provided by Tsiatis and Davidian (2004). Those authors note that joint modeling procedures can be classified in one of two ways, depending on the model used for the underlying longitudinal process with which failures are associated: the majority of papers assume that

true process follows a completely smooth trajectory described by usually only two random effects, whereas some papers (e.g. Wang and Taylor, 2001; Xu and Zeger, 2001) allow the subject-specific trajectory to have some noise which may have prognostic capability for the failure time process. The incorporation of a separate random effect governing short term fluctuations is quite difficult computationally (the two papers above use MCMC to fit their respective models), and this is why most authors choose to avoid it. Tsiatis and Davidian (2004) note that in most joint models “the value of the smooth trend [of the covariate]... is the predominant feature associated with prognosis”, and although it is admittedly a simplification of reality, computational limitations dictate its use. Papers by Fieuws and Verbeke (2006) and Rizopoulos et al. (2009) respectively use a pairwise modeling approach and Laplace approximations to ease the computational burden inevitably created in joint modeling by the numerical integration over random effects.

The normality assumption regarding the covariate process of subjects at risk that we made in Section 4.2.2 is similarly an obvious simplification of reality, but it does allow for calculation of the distribution of the covariate process at arbitrary times, which allows estimation of each subject’s complete covariate trajectory. Some joint models, for example Tsiatis and Davidian (2001), entirely avoid estimation of random effects and so cannot estimate subject-specific paths at all. With subject-specific paths in hand, the Cox model parameters for the hazard of failure (a hazard depending on the actual underlying covariate process, rather than only an over-smoothed version of it) can then be estimated.

This chapter will proceed as follows. Section 5.2 will show how to calculate the conditional covariate distribution at arbitrary points in time. Section 5.3.1

will introduce notation and Section 5.3.2 will give the theory for estimating the survival model parameters, while Sections 5.3.3 and 5.3.4 will discuss baseline hazard and survival probability estimation respectively. Section 5.4 will show some simulation results for the Cox model parameters, and Section 5.5 will provide an analysis of cardiac data. Section 5.5 will compare the predictive accuracy of different models using both ROC curves and the method of Schoop et al. (2008), which estimates dynamic prediction error using an inverse-probability-of-censoring-weighted (IPCW) estimator. This estimator is introduced in Section 5.3.5. Finally, Section 5.6 will wrap up the chapter and present some possible future research directions. Large sample theory is provided in Appendix K.

## 5.2 Covariate distribution at arbitrary times

Chapter 4 considered estimation of parameters driving the transition density of a longitudinal process, e.g. a covariate process. Our interest is now in using these estimates to identify the conditional distribution of the covariates of an arbitrary subject at an arbitrary future time. For progress to be made here, we assume that the covariate process transition density (conditioned on at-risk status) follows a multivariate normal distribution, but this assumption has already been made to deal with missingness in Section 4.2.2, so no further generality is lost. As mentioned in Section 3.2.3, in order to find the covariate distribution at times with no visits,  $A_i(t)$  may no longer be included in  $\{H_i^{m,c}(t)\}$  unless it is continuously observed.

Assumptions (4.1) and (4.2) are the vehicles used for imputation (i.e. extrap-



olation) at future times. Interpolation would produce efficiency gains and create a greater breadth of possibilities for terms in the Cox model (such as interpolated slopes of covariate trajectories); unfortunately it cannot be performed without bias. Consider the problem of finding the distribution of  $X(t)$  given  $X(t-1)$  and  $X(t+1)$ . Even with no missing data,  $E(X(t)|X(t-1), X(t+1), T \geq t+1)$ , and the corresponding variance are unknown. What is known is  $E(X(t)|X(t-1), T \geq t)$ , but once information on survival past time  $t$  is included, the normality (or whatever parametric assumption happened to be in use) would be destroyed.

The situation is no better if a marginal version of (4.1) and (4.2) is assumed, i.e. without conditioning on  $T \geq t$ . Not only would this make intermittently missing data impossible to deal with in the convenient conditional normal formulation used in (4.5), but different inverse-probability-of-survival-weights would have to be approximated to multiply each term of (4.9). Furthermore, the distribution of  $\{X(t+1)|X(t-1), T \geq t+1\}$  is not known and hence no Brownian-bridge argument for interpolation of  $X(t)$  would be applicable.

## 5.3 Methodology

### 5.3.1 Notation and assumptions

First recall from Section 4.2.1 that  $\Upsilon_i = \min(C_i, T_i)$  and  $\xi_i(t) = I(C_i \geq t)I(T_i \geq t)$ . Consider the usual Cox model hazard for mortality,

$$E(dN_i^{d*}(t)|H_i^d(t), C_i \geq t) = E(dN_i^{d*}(t)|H_i^d(t)) = \exp(\kappa_0^T h_i^d(t)) d\Lambda_0^d(t), \quad (5.1)$$

where  $dN_i^{d*}(t) = N_i^{d*}(t^- + dt) - N_i^{d*}(t^-)$ ,  $N_i^{d*}(t) = I(T_i \leq t)$ ,  $\{H_i^d(t)\} = \{X_i(t), \bar{X}_i(t^-), \bar{A}_i(t^-), T_i \geq t\}$ . The first equality in (5.1) is an independent censoring assumption. The vector  $h_i^d(t)$  is a realization of  $\{H_i^d(t)\}$ ; note in particular that we must assume  $h_i^d(t)$  cannot include previous unobserved values of the underlying covariate process because, for example, the distribution of  $\{X_i(t-k)|X_i(t), T_i \geq t\}$  is unknown when  $dN_i(t-k) = 0$ . Similarly,  $h_i^d(t)$  cannot include  $A_i(u)$ ,  $u \leq t$ , for  $u$  such that  $dN_i(u) = 0$  because its distribution is unknown, even at time  $t$ .

For use in this chapter, let  $N_i^d(t) = N_i^{d*}(\min(t, \Upsilon_i))$  denote the counting process of the observed failure for subject  $i$ ,  $\{\bar{H}_i^d(t)\}$  denote the observed portion of  $\{H_i^d(t)\}$ , which we assume includes  $\{T_i \geq t\}$ , with  $\{\bar{H}^d(t)\} = \bigcup_{i=1}^n \{\bar{H}_i^d(t)\}$ , and replace (4.3) and (4.4) with  $\tilde{\mu}_i(t) \equiv E(h_i^d(t)|\bar{H}_i^d(t))$  and  $\tilde{\Sigma}_i(t) \equiv \text{Var}(h_i^d(t)|\bar{H}_i^d(t))$ , the difference being that now  $A_i$  or a continuously observed  $A_i(t)$  can be included in  $\tilde{\mu}_i(t)$  and  $\tilde{\Sigma}_i(t)$ . Again, the dependence on  $\theta$  has been suppressed to avoid complexity of the notation, and will be done similarly for the rest of the chapter, but the reader should note that unless otherwise stated, the terms presented here depend on  $\hat{\theta}$ , and not the true data-generating parameter,  $\theta_0$ . Also note that  $h_i^d(t)$  conditional on  $\{\bar{H}_i^d(t)\}$  has a multivariate normal distribution, and that the methodology discussed in Section 5.2 provides the ability to estimate this distribution.

Finally, define  $\text{mgf}_i(t, \kappa) \equiv E(\exp(\kappa^T h_i^d(t))|\bar{H}_i^d(t)) = \exp(\kappa^T \tilde{\mu}_i(t) + \frac{1}{2}\kappa^T \tilde{\Sigma}_i(t)\kappa)$ , and notice that  $\text{mgf}_i(t, \kappa)$  is the moment generating function of  $h_i^d(t)$ , conditional on  $\bar{H}_i^d(t)$ , evaluated at  $\kappa$ .

### 5.3.2 Failure time process model

If the covariate process is continuously observed, the estimating equation for  $\kappa$  in the Cox model is (Kalbfleisch and Prentice, 2002):

$$\sum_{i=1}^n \int_0^\tau \left( h_i^d(t) - \frac{\sum_{j=1}^n \xi_j(t) h_j^d(t) \exp(\kappa^T h_j^d(t))}{\sum_{j=1}^n \xi_j(t) \exp(\kappa^T h_j^d(t))} \right) dN_i^d(t). \quad (5.2)$$

Taking the conditional expectation of (5.2) with respect to the observed covariate process would seem to be the desired solution for parameter estimation when the covariates are not continuously observed, but it is easily shown that this leads to a biased estimate of  $\kappa_0$ .

Tsiatis et al. (1995) provide one alternative solution in the similar context where covariates are observed with Gaussian measurement error. They assume that the covariate process of the subjects at risk remains normal for all times (as we do in (4.1) and (4.2)), and they replace  $\exp(\kappa^T h_i^d(t))$  with  $E(\exp(\kappa^T h_i^d(t)) | \bar{H}^d(t))$  in the usual Cox model likelihood. This leads to the following estimating equation for  $\kappa$ :

$$\sum_{i=1}^n \int_0^\tau \left[ E(h_i^d(t) | \bar{H}_i^d(t)) - \frac{E(\sum_{j=1}^n \xi_j(t) h_j^d(t) \exp(\kappa^T h_j^d(t)) | \bar{H}^d(t))}{E(\sum_{j=1}^n \xi_j(t) \exp(\kappa^T h_j^d(t)) | \bar{H}^d(t))} \right] dN_i^d(t). \quad (5.3)$$

By switching the order of the expectation and derivative, all the expectations in (5.3) are easily found. Since (5.3) amounts to using the conditional expectation of the hazard in (5.1), instead of the hazard itself, in the partial likelihood, it leads to biased estimates of  $\kappa_0$ . But as Dafni and Tsiatis (1998) show, the bias caused by (5.3) is greatly reduced from more naïve methods like last value carried forward. Another problem, as discussed in Section 4.1, is that the assumption of normality conditioned on at-risk status may not even lead to processes that exist. The conditional normality is however a very convenient mathematical assumption that facilitates computation. Marginal Gaussian assumptions

are quite common in the longitudinal data literature (including Chapter 3) and as seen in Section 4.3, (4.1) and (4.2) are only negligibly different than (3.8) and (3.9), so the conditional normality is not unreasonable to assume.

Our proposed estimating equation for  $\kappa$  is quite similar to (5.3). It is:

$$U^\dagger(\kappa, \hat{\theta}) = \sum_{i=1}^n \int_0^\tau \left[ E(h_i^d(t) | \bar{H}_i^d(t)) - \frac{\sum_{j=1}^n \xi_j(t) E(h_j^d(t) | \bar{H}^d(t)) E(\exp(\kappa^T h_j^d(t)) | \bar{H}^d(t))}{\sum_{j=1}^n \xi_j(t) E(\exp(\kappa^T h_j^d(t)) | \bar{H}^d(t))} \right] dN_i^d(t). \quad (5.4)$$

This can be viewed as a special case of the Expectation-Substitution (ES) algorithm (Elashoff and Ryan, 2004), which in its full generality alternates between replacing functions of complete data with their expected values and substituting these expected values into a complete data estimating equation, which is then solved. Equation (5.4) is simply their S-step run once. No iteration between E and S steps is required because the expectations in (5.4) use  $\hat{\theta}$  as previously estimated in (4.9) rather than simultaneously estimating  $\theta_0$  and  $\kappa_0$ .

In exchange for our strong assumption on the distribution of the longitudinal data, we do not have to make the strong assumption made in the literature on joint modeling of longitudinal and time-to-event data (Tsiatis and Davidian, 2004) that the failures only depend on the smoothed approximation of the covariate process. This is quite evident from the fact that (5.4) includes  $h^d(t)$ .

Each contribution to the estimating equation for the Cox model with continuously observed data, (5.2), is the covariates of the subject who failed minus a weighted average of the covariates of all the subjects still at risk with weights proportional to the hazards of failure. An intuitive explanation for (5.4) is that it is the analog in the discretely observed data setting: the expected value of

the covariates of the subject who failed minus the weighted average of the expectations of the subjects still at risk with weights proportional to the expected hazards of failure. For (5.4) to lead to a consistent estimate of  $\kappa_0$ , we do require the covariates to remain normal conditioned on at-risk status.

An alternative EE would also involve conditioning on  $T_i = t$  in each piece of (5.4). This would arguably be the analog of (5.2) for discretely observed data. The required expectations would still be available in closed form because a simple Bayes' rule calculation shows that  $h_i^d(t)$  is conditionally multivariate normal with

$$\begin{aligned} E(h_i^d(t)|\bar{H}_i^d(t), T_i = t) &= \tilde{\mu}_i(t) + \tilde{\Sigma}_i(t)\kappa_0, \\ \text{Var}(h_i^d(t)|\bar{H}_i^d(t), T_i = t) &= \tilde{\Sigma}_i(t). \end{aligned} \tag{5.5}$$

However, as the following theorem shows, (5.4) is appropriate for unbiasedness and consistency.

**Theorem 5.3.1.** Given the conditions for Theorem 4.2.1, given assumption (5.1), and given conditions (K1)-(K5) in Appendix K.1, the  $\hat{\kappa}$  that solves  $0 = U^\dagger(\hat{\kappa}, \hat{\theta})$  is a consistent and asymptotically normal estimator of  $\kappa_0$ , the true parameter.

The proof of Theorem 5.3.1 is provided in Appendix K.2, along with an estimate of the asymptotic variance of  $\hat{\kappa}$ .

### 5.3.3 Baseline hazard estimation

The cumulative baseline hazard must be estimated in order to predict survival for arbitrary subjects after (4.9) and (5.4) (or (5.3)) have been solved. When

continuously observed data is available, the Breslow estimator is used. It is:

$$\hat{\Lambda}_0^d(t) = \int_0^t \sum_{i=1}^n \frac{1}{\sum_{j=1}^n \xi_j(s) \exp(\hat{\kappa}^T h_j^d(s))} dN_i^d(s). \quad (5.6)$$

With intermittently observed data, our proposed estimator is different, but it is the same for both (5.3) and (5.4):

$$\hat{\Lambda}_0^d(t) = \int_0^t \sum_{i=1}^n \frac{1}{\sum_{j=1}^n \xi_j(s) E\left(\exp(\hat{\kappa}^T h_j^d(s)) | \bar{H}^d(s)\right)} dN_i^d(s). \quad (5.7)$$

### 5.3.4 Probability of survival

With continuously observed data, the survivor function for subject  $i$  is usually estimated with (Kalbfleisch and Prentice, 2002):

$$\hat{S}_i(s, s+t) = \exp\left(-\int_s^{s+t} \exp(\kappa^T h^d(u)) d\Lambda_0^d(u)\right), \quad (5.8)$$

where  $\Lambda_0^d(u)$  is replaced by an estimate. With intermittent observations, the conditional expectation of (5.8) is not available in closed form, but the expectation of its logarithm is, and (for both (5.3) and (5.4)) this expectation leads to

$$\hat{S}_i^{(s)}(s, s+t) = \exp\left(-\int_s^{s+t} \text{mgf}_i^{(s)}(u, \hat{\kappa}) d\hat{\Lambda}_0^d(u)\right), \quad (5.9)$$

$$\hat{S}_i^{(s+t)}(s, s+t) = \exp\left(-\int_s^{s+t} \text{mgf}_i^{(s+t)}(u, \hat{\kappa}) d\hat{\Lambda}_0^d(u)\right), \quad (5.10)$$

where  $\text{mgf}_i^{(b)}(t, \kappa)$  is the same as  $\text{mgf}_i(t, \kappa)$  except  $\tilde{\mu}_i(t)$  and  $\tilde{\Sigma}_i(t)$  are based on information observed at or before time  $b$  (rather than time  $t$ ). Equation (5.9) gives an estimate of the probability of survival to time  $s+t$  given survival to time  $s$ , whereas (5.10) conditions on the observed covariate path between  $s$  and  $s+t$  to give the probability that a future subject following this covariate path would survive to time  $s+t$  having already survived to time  $s$ . Unfortunately, due to conditioning on survival past  $u$ , the distribution of  $X_i(u)$  cannot be calculated

exactly for  $s < u < s + t$ . Nonetheless, interpolation could be useful in evaluating prediction error, as will be explained now.

### 5.3.5 Evaluating prediction error

With survival estimates now obtained, a way to measure the prognostic capabilities of the candidate models must be defined. One useful and convenient metric is the quadratic loss function (Schoop et al., 2008):

$$\left(I(T_i > s + t) - \hat{S}^{(s)}(i, s, s + t)\right)^2, \quad (5.11)$$

which is defined for each  $i$  and for some choices of  $s$  and  $t$ . Defining  $\hat{S}^{(s)}(i, s, s + t)$  to be constant at 0.5 gives 0.25 as the quadratic loss, so any sensible prediction scheme has 0.25 as an upper bound for its expected quadratic loss.

Gerds and Schumacher (2006, 2007) study this model assessment problem with time-fixed covariates ( $s = 0$ ) by using an inverse-probability-of-censoring-weighted (IPCW) estimator of the prediction error (PE), where PE is defined as the marginal expectation of (5.11). Schoop (2008) and Schoop et al. (2008) extend the work of Gerds and Schumacher to dynamic predictions involving possibly internal time-dependent covariates, defining PE as the expected value of (5.11) conditional on survival to time  $s$ , and selecting the best model to be the one with the lowest PE estimate. They remark that by varying either  $s$  or  $t$  one can get a prediction error curve, and they define time-averaged prediction error as an integral of the PE over the possible values of  $s$  or  $t$ . For computational convenience, we propose randomly sampling  $s$  independently of  $\hat{S}$ , and using  $s = s_1, s_2, \dots, s_m$  for each subject, keeping  $m$  small. Similarly to Schoop et al. (2008), the resulting consistent estimate of the true time-averaged prediction

error for a fixed  $t$  is

$$\widehat{\text{PE}}(s, t) = \frac{1}{nm} \sum_{i=1}^n \sum_{j=1}^m \frac{\xi_i(s_j) \left( I(T_i > s_j + t) - \hat{S}^{(s_j)}(i, s_j, s_j + t) \right)^2 W_i(s_j, s_j + t)}{\hat{P}(T_i > s_j) \hat{P}(C_i > s_j)}, \quad (5.12)$$

where

$$W_i(s, s + t) = \frac{\xi_i(s + t)}{\hat{P}(C_i > s + t | \xi_i(s) = 1, H^c(s))} + \frac{\delta_i(1 - \xi_i(s + t))}{\hat{P}(C_i > \Upsilon_i | \xi_i(s) = 1, H^c(s))},$$

and  $\{H^c(t)\}$  is the information available at time  $t$  to predict  $dN^c(t) \equiv N^c(t^- + dt) - N^c(t^-)$ , the indicator of a censoring event at time  $t$ . The denominators in  $W_i$  are estimated using an equation analogous to (5.9) and by assuming coarsening at random, (4.6).

For the second term in the denominator of (5.12), Schoop et al. (2008) define

$$\hat{P}(C_i > s_j) = \prod_{k=1}^{r_i(s_j)} \hat{P}(C_i > t_{ik} | H^c(t_{ik})) \cdot \hat{P}(C_i > s_j | H^c(s_j)),$$

where  $r_i(t)$  is the time of the most recent visit before time  $t$ . They suggest estimating  $\hat{P}(T_i > s_j)$  using the Kaplan-Meier estimator, noting that this makes the IPCW estimator's weighting scheme independent of the survival model in question, implying that it remains unbiased even under misspecification of the survival model.

Our model suggests that a possible alternative term is (5.10) evaluated with  $s = 0$  and  $t = s_j$ . However, if this is done, the IPCW estimator's weighting scheme is no longer independent of the survival model, resulting in its bias if the survival model is misspecified (there is also the small bias discussed in Section 5.3.4). But as described in Schoop (2008), the intuitive reasoning for inverse-weighting at time  $t$  is that a subject who only had a probability  $p$  of remaining at risk until time  $t$  has to represent  $p^{-1}$  other similar subjects (i.e. ones who had the same covariate path) who ceased to be at risk sometime before  $t$ .



## 5.4 Simulations

This chapter introduced the estimating equation (5.4), and we now want to study the properties of the estimates it produces for  $\kappa_0$ . See Sections 3.3 and 4.3 for a complete description of the setup. Interest has now shifted away from the estimation of the longitudinal process, but it must still be estimated before estimating  $\kappa_0$ . We will not report results of its estimation here though because that was done in Section 4.3. The reader is reminded that the longitudinal processes were simulated to be marginally normal, that is, without any conditioning on  $T \geq t$ , so (5.4) is subject to misspecification in these simulations. It turns out that (5.4) and (5.3) produce virtually the exact same estimates in all cases, so we will not report both.

The terminal events were simulated according to (5.1) with  $\kappa_0 = (0.3, 0.2, 0.05, 1)$ , constant baseline hazard, and  $h^d(t) = (X_1(t), X_2(t), Z_1, Z_2)$ .

Table 5.1 gives results for  $\mu(t, \beta) = \beta_0 - 0.5t + 0.03t^2$  using (5.4), and Table 5.2 gives results for the mean reverting process using (5.4). With only one failure per subject rather than about 10 visits per subject (the information available to estimate parameters in Sections 3.3 and 4.3), the relative biases are higher and the ASE estimation is unreliable, hence it is not reported. Simulations with higher  $n$  were considered, and they did show ASE comparable to ESE, but to undertake such simulations for all combinations of mean function and censoring distribution would have been computationally prohibitive. Interestingly, the simulations with more failures occurring (due to a higher baseline intensity), tended to have higher biases despite their lower ESE. Some of that behavior may be due to having 20% fewer visits when censoring was 50% as opposed to

Table 5.1: Simulation results for deterministic drift using (5.4) with 400 subjects and both 50% censoring and 75% censoring

	50% censoring	75% censoring
Parameter	rBias (ESE)	rBias (ESE)
$\kappa_1$	-0.003 (0.065)	0.003 (0.067)
$\kappa_2$	-0.069 (0.064)	-0.020 (0.073)
$\kappa_3$	0.032 (0.006)	0.004 (0.010)
$\kappa_4$	-0.002 (0.173)	-0.032 (0.204)

Table 5.2: Simulation results for mean reverting drift using (5.4) with 400 subjects and both 50% censoring and 75% censoring

	50% censoring	75% censoring
Parameter	rBias (ESE)	rBias (ESE)
$\kappa_1$	0.038 (0.206)	0.000 (0.265)
$\kappa_2$	-0.091 (0.175)	-0.041 (0.267)
$\kappa_3$	-0.010 (0.008)	0.014 (0.011)
$\kappa_4$	0.032 (0.167)	0.013 (0.237)

75%.

## 5.5 Data Analysis: Cardiac care unit

These data were kindly provided by Dr. Mark Cowen of the Quality Institute, part of the St. Joseph Mercy Health System in Ann Arbor, Michigan. The complete dataset consists of 23,792 unique subjects with a total of about 475,000 observation times; the analyses reported below use approximately one quarter

of this complete dataset. Upon admission to the hospital, time-fixed covariates are obtained for every subject. At observation times, some subset of the time-dependent covariates is measured; importantly, this subset may be different for different observation times. As an exploratory analysis, we fit some basic logistic models to discover which time-dependent covariates correlated highly with mortality time, with the intention of only modeling the trajectories of these selected covariates. Table 5.3 provides definitions of the six resulting possible time-dependent covariates; these covariates average approximately 40% missingness across all subjects. They are skewed, so we used a log transform of all of them.

Table 5.3: Time-dependent covariate definitions

Abbreviation	Definition
cre	serum creatinine
hgb	hemoglobin
mag	serum magnesium
pot	serum potassium
sod	serum sodium
wbc	white blood count

The raw data include some negative test times which are due to subjects undergoing tests in the intensive care unit before arrival into the cardiac unit at time zero. In addition, time zero generally corresponds to different times of day for different subjects, so before applying (4.9) and (5.4) to the data, we defined time zero for each subject as 12:00AM on the day of their first test. Figure 5.1 shows the resulting baseline cumulative hazard estimate over the first 60 hours when (3.6) is fit with only the covariates from Tables 5.3 and 5.4 as  $h^{v,m}$ . As

expected, this shows a clear spike in testing at approximately 5:00-7:00AM each day, with a lower spike on the first day because many subjects are admitted after 7:00AM. This spike is due to routine testing, hence tests taken at this time of day are not as informative for covariate values, and the proportional hazards assumption of the Cox model for visitation is violated. Therefore, if one is fitting a marginal model (Section 3.2.2), (3.6) should be fit using the time-fixed covariates as well as the log of the time-dependent covariates and their interaction with an indicator of  $5 < t \bmod 24 < 7$  as " $h^{v,m}$ ". If one is fitting a conditional model (Section 3.2.3), as we are in this section, (3.6) is not utilized and the only consideration is whether  $X(t)$  depends on  $t \bmod 24$  even after conditioning on  $\bar{X}(t-)$  and  $\bar{A}(t-)$ . If it does, then  $t \bmod 24$  must be included in  $\{H^{m,c}(t)\}$  as well as in  $\{H^{v,c}(t)\}$  (see (4.8)). We notice no such dependence, so  $\{H^{m,c}(t)\}$  only includes the time-fixed covariates and the previous time-dependent covariates.

Table 5.4: Time-fixed covariate definitions

Abbreviation	Definition
age	age
ffp	fresh frozen plasma transfusion indicator
ps	potassium supplementation indicator
rbc	red blood cell transfusion indicator

The mean reverting model does not fit these data well, and there is no clear trend in covariates over time, so we assumed a zero drift process with unspecified variance:

$$(X(s+t)|X(s)) \sim N(X(s), \log(1+t) \cdot \Sigma),$$

where  $X$  consisted of a subset of the covariates from Table 5.3. The possible choices for " $h^d(t)$ " are listed in Table 5.5. See Table 5.4 for definitions of the

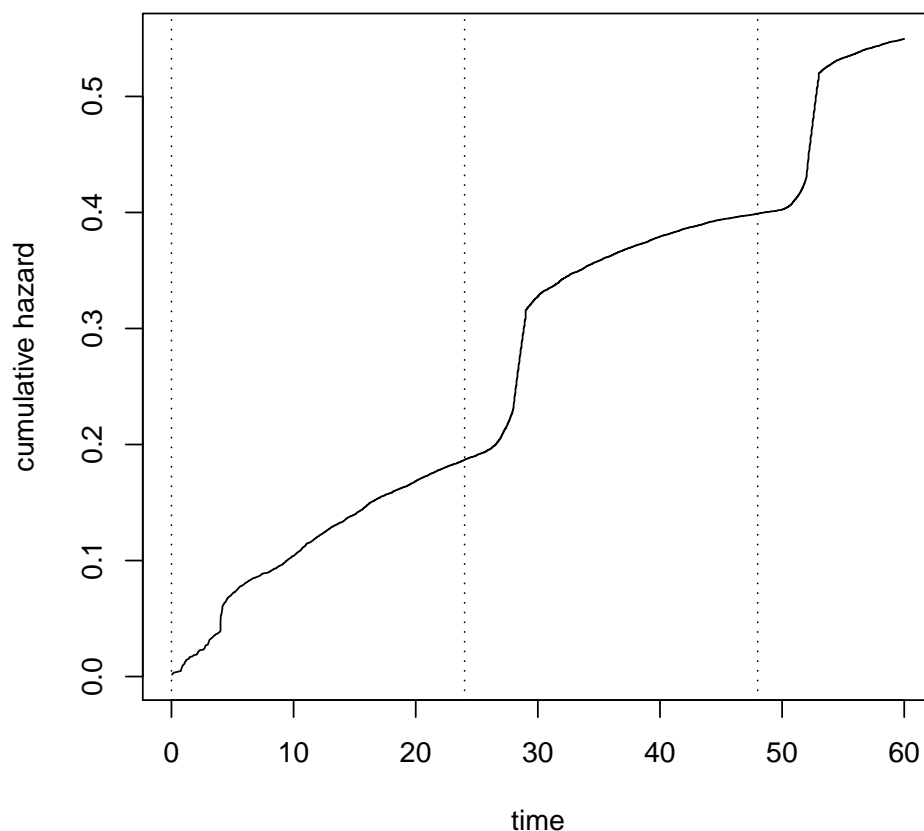


Figure 5.1: Baseline cumulative hazard of test times in the first 60 hours. The dotted vertical lines denote midnight. This model uses all the covariates from Tables 5.3 and 5.4 in the Cox model for visitation

time-fixed covariates.

Estimating equations (4.9) and (5.4) were solved for each of these models; actually in order to estimate prediction error, we partitioned the data into three smaller pieces, and fit each model to each partition, and then validated each training set with the data from the other two pieces (see Gerds and Schumacher (2007) for a description of similar 3-fold cross-validation).

We used a 12 hour window as a time-frame for prediction. According to Dr. Cowen, this is short enough that patients could fall through the cracks on a shift change, for example, but long enough so that a diversion of resources could still help the patient. Equation (5.12) was used to estimate prediction error starting from randomly sampled time points, and the average of the three training sets for each model is presented in Table 5.5. The corresponding ROC curves are displayed in Figure 5.2. Each survival probability was calculated using (5.9) and if it was greater than  $C$ , survival was predicted. The predicted survival probability was never less than 0.98, so  $C$  ranged from 0.98 up to 1.

Table 5.5: Prediction error estimates for different possible models

Model #	covariates used	$\widehat{PE}$
1	cre,hgb,mag,pot,sod,wbc,ffp,age,ps,rbc	$1.638 \times 10^{-2}$
2	cre,hgb,mag,pot,wbc,ffp,age,ps	$1.637 \times 10^{-2}$
3	cre,mag,pot,wbc,ffp,age,ps	$1.611 \times 10^{-2}$
4	cre,pot,wbc,ffp,ps	$1.609 \times 10^{-2}$
5	cre,wbc,ps	$1.616 \times 10^{-2}$

Table 5.5 doesn't provide much to choose between the models, but these prediction error point estimates are noticeably lower than the PE estimates from models not including important covariates like cre and wbc. The best model appears to contain cre, pot, wbc, ffp, and ps, and by using

$$h^d(t) = (\text{cre}(t), \text{pot}(t), \text{wbc}(t), \text{ffp}, \text{ps}),$$

the average of the three estimates of  $\kappa_0$  is:

$$\hat{\kappa} = \begin{pmatrix} 0.54(0.17) & -0.03(0.15) & 0.59(0.24) & 1.42(0.42) & 0.21(0.28) \end{pmatrix},$$

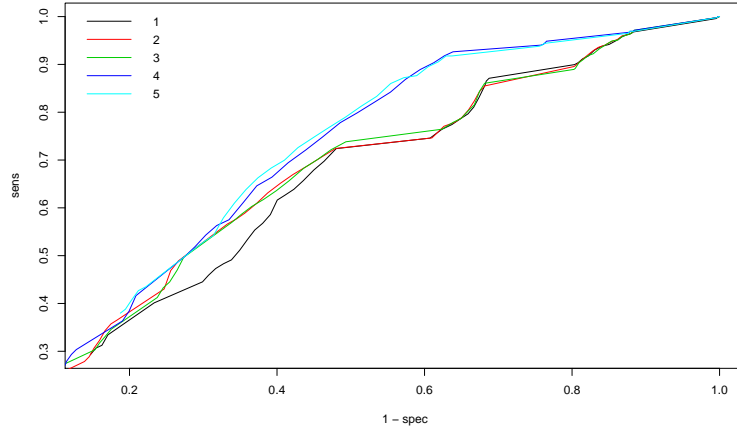


Figure 5.2: ROC curves plotting sensitivity versus 1-specificity for different possible models

where the ASE is in parentheses. About 97% of patients are censored, so there is not much information to estimate  $\kappa_0$ , as seen from the large ASE for the preceding estimates. The reported ASE is the average of the three ASEs, which are estimated separately as the square root of  $n^{-1}$  times the diagonal of an estimate of (K.15) (see Appendix K.2). Using cre, pot, and wbc, the average of the three estimates led to

$$\hat{\Sigma} = \begin{pmatrix} 0.009 & 0.002 & 0.003 \\ 0.002 & 0.008 & 0.001 \\ 0.003 & 0.001 & 0.028 \end{pmatrix},$$

where each estimate had a negligible asymptotic standard error. We chose  $\log(1 + t) \cdot \Sigma$  instead of the Brownian motion-induced  $t \cdot \Sigma$  because the latter leads to some extremely influential tests when covariates hardly change over a long gap time between visits.

Figure 5.3 shows plots of standardized residuals (for symmetry the mag residuals have been included too, but the model with just cre, pot, and wbc gives virtually the same residuals) for each of the time-dependent covariates.

Each residual is the difference between the observed (or projected for those that were unobserved) covariate and the mean of the transition density all divided by the square root of the variance of the transition density. Obviously there are several high residuals, but QQ plots show that the tails of these distributions are quite similar to those of a  $t$ -distribution with 4 degrees of freedom, and our simulations in Sections 3.3 and 4.3 showed that the bias created in the presence of heavy tailed errors tends to be minimal.

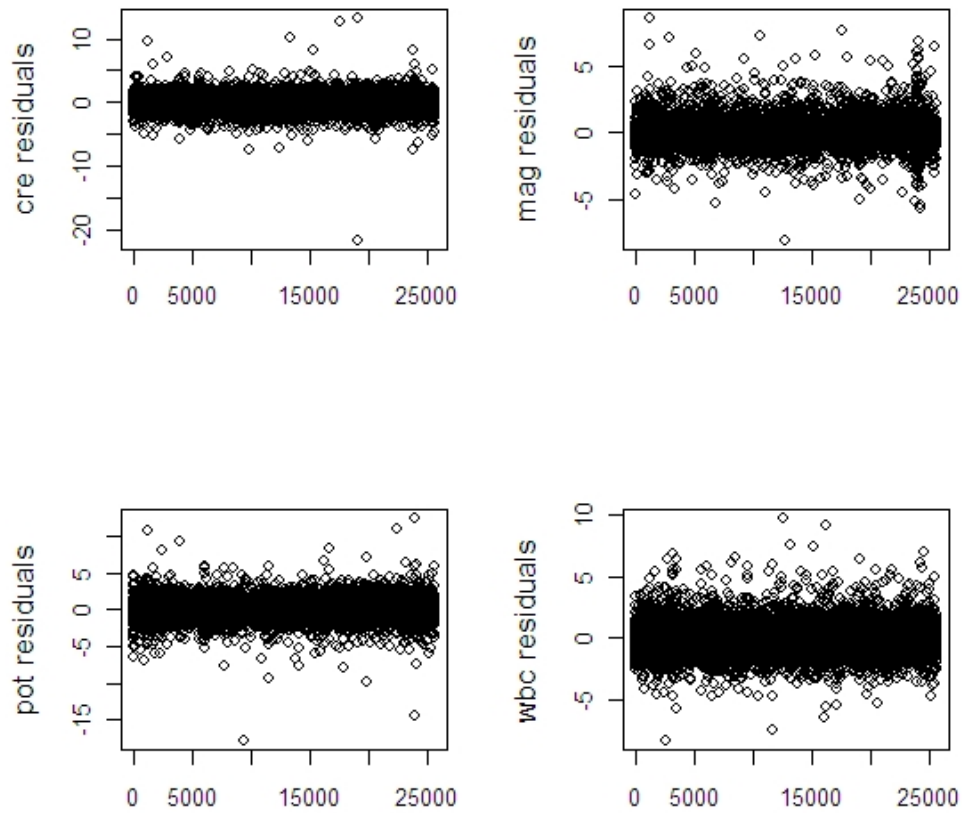


Figure 5.3: Standardized residuals for covariates in the cardiac data



## 5.6 Discussion

This chapter has introduced an alternative to the commonly used strategy of jointly modeling longitudinal and time-to-event data when the ultimate goal is to model the survival process. The majority of the joint modeling literature assumes that the event time process cannot depend on short term fluctuations in the longitudinal (covariate) process; this assumption is questionable, so we replace it with an assumption of conditional normality of covariates among those subjects still at risk. While the latter may create processes that do not technically exist (Tsiatis et al., 1995), it is mathematically convenient in terms of facilitating computation, and it allows the event time process to depend on short term fluctuations in the covariate process.

We have proposed an alternative to the estimating equation presented in Tsiatis et al. (1995), and shown consistency and asymptotic normality of our parameter estimates under the aforementioned conditional normality assumption.

Simulations have shown that  $\kappa_0$  can be estimated with relative bias of usually at most only 4% when only 100/400 subjects experience a failure, and an analysis of the cardiac care unit dataset has selected appropriate time-fixed and time-dependent covariates to use in prediction of future deaths. The prediction model incorporating these covariates will hopefully enable hospital resources to be assigned to patients in the greatest immediate need.

If one wants to avoid the tenuous conditional normality assumptions, with mean and variance specified by (4.1) and (4.2) respectively, then the usual marginal assumption could be offered instead - that is, no conditioning on survival in the Gaussian assumption. While more technically sound, this prohibits

the simple handling of missing data proposed in chapter 4; it also prohibits simple extrapolation provided in this chapter. An inverse-probability-of-survival-weight (IPSW) might be feasible in the complete data setting, and further investigation might show this to be a reasonable alternative to the previously discussed smoothing assumptions made in the joint modeling literature, while still providing more technically sound assumptions for large sample theory; it would however destroy the simplicity of our proposed methods.

Another direction for future study could be to model a discrete response conditional on its past observations using a continuous-time Markov chain (CTMC), handling missing data with a multinomial imputation method. The CTMC could be used to model the length of time spent in various states, and to predict the time-to-event of an associated failure process. A possible application of this would be using credit ratings in financial data to predict bankruptcy.

## CHAPTER 6

### CONCLUSION

This thesis has studied longitudinal data in several different forms. In Chapter 2 the focus was on modeling the times between measurements, in Chapters 3 and 4 interest was in modeling the measurement process itself, and in Chapter 5 the longitudinal measurements were used to predict a time-to-event of an associated process.

Chapter 2 began by introducing the methods for analyzing recurrent events, and noted that they can generally be cross-classified into one of four categories determined by: (i) the choice of “calendar” versus “gap” time as the fundamental temporal scale; and, (ii) the use of “marginal” versus “intensity” models for analyzing the data. The main advantage enjoyed by marginal models over intensity models is that there is no need to fully specify the within subject covariance structure. This makes them useful for studying population parameters, but not as useful for prediction. Intensity models, due to their full specification of covariance structure, can be subject to misspecification, but they do allow for a weaker independent censoring assumption because they condition on event histories as well as covariate histories. The gap time approach in Chapter 2, which builds upon the work of Murphy et al. (1995), uses a conditional generalized estimating equation (condGEE; also see the R package of the same name) that relaxes the stringent restrictions imposed by simpler marginal models while avoiding the need to fully specify how the probability of subsequent recurrence depends on the prior event and covariate histories.

As opposed to modeling the less interpretable event intensities, the condGEE procedure directly models the first two moments of the gap times conditional

upon the previous event history; this conditional specification differs from GEE as proposed by Liang and Zeger (1986) and GEE2 proposed by Prentice and Zhao (1991). The conditional structure creates an interesting class of possible choices for the working variance, and as we show in Appendix C, upper triangular choices are required for an unbiased estimating equation.

We dealt with the censored gap time by using a parametric assumption (which simulations showed to be quite robust) on its distribution, but further research could consider different options such as inverse-probability-of-censoring-weighting (IPCW), which would ignore the censored gap time and up-weight the contributions of the complete gap times according to their probability of having been censored. This would require specification of a censoring model rather than a model for the length of the censored gap time for each subject. It would also be interesting to consider multivariate recurrent event processes arising either as a result of clustering or due to the presence of multiple recurrent event outcomes on each subject. Modeling the dependence structure would likely cause the robustness of the present approach to suffer and it may be advantageous to use a proper extension of IPCW-type estimation.

Chapter 3 departed from the recurrent event setting; parameters governing the intensity of measurement (i.e. intensity of visitation) were considered a nuisance, but they are still accounted for to avoid bias in the estimation of the parameters describing the evolution of the longitudinal process. Papers by Lin et al. (2004) and Bůžková and Lumley (2009) studied this class of problems by employing estimating equations, with the ground-breaking contribution made by Lin and Ying (2001). The former two papers use an inverse-intensity-weight

to allow for visitation dependent on previous responses. The work in this thesis expands on theirs by allowing for the possibility of modeling the response process not only conditional upon covariates but also upon its own history (that is, a conditional estimating equation rather than a marginal estimating equation), and also by handling a multivariate process, which of course requires variance estimation as well.

We explored the differences between marginal and conditional estimating equations: these differences extend to the types of problems they can answer, and the assumptions required, particularly on the visitation process. It turns out that with conditional estimating equations (CEEs), no inverse-intensity-of-visitation-weight is required for consistency of the estimates of the parameters driving the longitudinal process.

Chapter 4 introduced the concept of intermittent missingness in response process: that is, dimensions which may be missing for a particular visit only to be measured again at a future visit. Properly dealing with intermittent missingness of a multivariate response in longitudinal data is an area that has not received much attention in the literature, but by using CEEs to estimate transition density parameters of the multivariate process, and hence describe its evolution, the missingness can be handled relatively easily with a sequentially missing at random (S-MAR) assumption. This does require a parametric assumption (on the transition density of the longitudinal process given at-risk status) to find the expectation of the unobserved data given the observed data in order to maintain an unbiased estimating equation, but simulations showed a robustness to this assumption. We used a Gaussian assumption, but in future work it might be interesting to also consider distributions besides the Gaussian,

with the possible use of copulas to model the dependence. This chapter also considered estimation of the longitudinal process in the presence of a terminal event. This was done by simply modeling the mean and variance conditional on survival; the conditional assumptions required for this parameter estimation are complicated by the parametric assumption used to handle missingness, and assuming that a process is normal conditioned on at-risk status for all times is theoretically questionable, but practically quite reasonable (Tsiatis et al., 1995). Misspecified simulations showed that the parameter estimates conditional on at-risk status are virtually the same as the unconditional parameters.

The need for the parametric transition density precludes a similar approach to marginal modeling with missing data (the longitudinal process history is part of the transition density), but the CEE approach could be used to model a time-dependent covariate process subject to missingness, which could then be used in a marginal estimating equation for a response without missingness.

Finally, Chapter 5 used the estimated trajectories of the longitudinal process to predict survival. The study of this topic was motivated by a dataset of cardiac patients from Dr. Mark Cowen in Ann Arbor, Michigan. Each patient has time-fixed covariates measured upon entry and time-dependent covariates measured intermittently throughout their stay, but the time-dependent covariate measurements are subject to missingness. We broke the problem into two steps: (1) estimate the underlying covariate process (using methodology from Chapter 4), and (2) use this process to predict future survival of arbitrary patients, specifically, based on covariate measurements up to time  $s$  predict survival to time  $s + t$ . This type of prediction has been termed both partly conditional hazard modeling (Zheng and Heagerty, 2005) and landmark analysis (Van Houwelin-

gen, 2007).

Before prediction takes place, parameters driving the failure time process are estimated by maximizing a partial likelihood (PL) based on that of the well-known Cox model, the difference being that the intermittent observation scheme creates some conditional expectations in the PL. The pointwise knowledge of the covariate trajectory distribution (available from the Gaussian assumption made in Chapter 4 to deal with missingness) can allow the failure hazard to depend on a function that can be estimated from the history of the observed covariates, for example the slope of the covariate process at times after  $s$ . The estimates derived from the PL model are then used to create a prediction model. The models created from different sets of predictors can then be compared in the usual ways: with ROC curves or by using a method similar to Schoop et al. (2008), who estimate dynamic prediction error using an IPCW estimator.

Our approach to step 2 differs from many in the literature (Tsiatis and Davidian, 2004 provide a good review) because we do not jointly model the longitudinal and time-to-event processes by connecting them with an unobserved latent variable. The downside of our methodology is our assumption that conditional on at-risk status, the covariates for each subject follow a normal transition density. But the upside is the ability to consider step 1 on its own, and the ability to model subject-specific trajectories; Tsiatis and Davidian (2004) note that in most joint models “the value of the smooth trend [of the covariate]... is the predominant feature associated with prognosis”, and that the incorporation of a separate random effect governing short term fluctuations is quite difficult computationally, and is usually avoided. Our conditional methodology provides

the same upside as including this random effect, but without the severe computational challenges of numerical integration, and hence, unlike the majority of joint modeling literature, allows failures to depend on short term fluctuations in health, rather than only on a smooth trend which is often only described by two random effects.

We have laid out assumptions connecting the visitation, covariate, failure time, and censoring processes, but there is work to be done in weakening these assumptions, especially in the missing data context. Some papers, for example Sun and Tong (2009), have begun to do this in the full data context with the use of latent variables.

A direction for future study could be to model a discrete response conditional on its past observations using a continuous-time Markov chain (CTMC), handling missing data with a multinomial imputation method. The CTMC could be used to model the length of time spent in various states, and to predict the time-to-event of an associated failure process. A possible application of this would be using credit ratings in financial data to predict bankruptcy.



## APPENDIX A

### REGULARITY CONDITIONS (CHAPTER 2)

Regularity conditions sufficient for Theorems 2.2.1, 2.2.2 and 2.2.3 to hold are summarized below:

- (A0) The parameter  $\eta = (\theta^T, \sigma)^T$  lies in some compact subset  $\mathcal{O} \subset \mathbb{R}^p$ ; the data generating parameter  $\eta_0$  is assumed to lie interior to  $\mathcal{O}$ . The known transformation  $h(x)$  is monotone nondecreasing and bounded for  $x \in (0, \infty)$ . Subjects are independent and identically distributed. Noninformative censoring holds, in the sense that for  $j \geq 1$ , we have  $E[Y_{1j}|H_{1j}] = E[Y_{1j}|\mathcal{H}_{1j}]$  and  $\text{Var}[Y_{1j}|H_{1j}] = \text{Var}[Y_{1j}|\mathcal{H}_{1j}]$  for  $H_{1j} = \mathcal{H}_{1j} \cup \{C_1 \geq S_{1,j-1}\}$ .
- (A1) Assumption (2.10) holds, with  $F_0(\cdot)$  correctly specified. In addition,  $b_{ij}(\eta)$ ,  $\mu_{ij}(\theta)$ , and  $V_{ij}^{-1}(\theta)$  are each bounded and twice continuously differentiable for  $i, j \geq 1$ ,  $\eta \in \mathcal{O}$ .
- (A2)  $S_n(\eta)$  is continuous for  $\eta \in \mathcal{O}$ , and  $S_n(\eta)$  converges uniformly in probability to  $S(\eta) := E_{\eta_0}[\psi(\eta, \mathbb{O}_1)]$  in some open neighborhood containing  $\eta_0$ , where  $S(\eta_0) = 0$ .
- (A3)  $S'_n(\eta) := \frac{d}{d\eta} S_n(\eta)$  exists and is continuous for  $\eta \in \mathcal{O}$ , and  $S'_n(\eta)$  converges uniformly in probability to  $S'(\eta) := \frac{d}{d\eta} S(\eta)$  in some open neighborhood containing  $\eta_0$ .
- (A4)  $S'(\eta_0)$  is non-singular.
- (A5)  $\psi(\eta, \mathbb{O}_1)$  satisfies the Lipschitz condition

$$\| \psi(\eta_1, \mathbb{O}_1) - \psi(\eta_2, \mathbb{O}_1) \| \leq \dot{\psi}(\mathbb{O}_1) \| \eta_1 - \eta_2 \|,$$

where  $\eta_1$  and  $\eta_2$  both lie in a neighborhood containing  $\eta_0$  and  $\dot{\psi}(\mathbb{O}_1)$  is a measurable, scalar-valued function with  $E_{\eta_0}(\dot{\psi}^2(\mathbb{O}_1)) < \infty$ .

$$(A6) \quad E_{\eta_0} \|\psi(\eta_0, \mathbb{O}_1)\|^2 < \infty.$$

$$(A7) \quad S_n(\widehat{\eta}_n) = o_p(n^{-1/2}).$$

Conditions (A0) and (A1) impose assumptions specific to the problem at hand; conditions (A2)-(A7) are more general, consisting of a combination of regularity conditions taken from Yuan and Jennrich (1998) and van der Vaart (1998, §5.3) adapted to the current problem. Yuan and Jennrich (1998), extending results originally due to Foutz (1977), use (A2)-(A4) and the inverse function theorem to prove consistency of a sequence of solutions obtained via an unbiased estimating equation. van der Vaart (1998, §5.3) uses (A4)-(A7) and the assumption that a consistent estimator exists in order to prove asymptotic normality under an i.i.d. sampling assumption. In particular, under conditions (A0)-(A7), the proofs of Theorem 2.2.2 and 2.2.3 are respectively direct consequences of Theorem 3 of Yuan and Jennrich (1998) and Theorem 5.21 in van der Vaart (1998). The remaining details are therefore omitted.

Condition (A1) says very little about the nature of  $K_r(\cdot)$ ,  $r = 1, 2$  appearing in (2.14) and (2.15); the requisite assumptions are embedded in (A2)-(A7). For example, the derivatives appearing in (A3)-(A5) involve the functions  $k_h(s) := \frac{d}{ds} K_h(s)$ ,  $h = 1, 2$ . Using integration by parts and assuming that  $F_0(\cdot)$  is continuously differentiable, we see that

$$k_h(s) = \lambda_0(s) [K_h(s) - s^h], \quad (A.1)$$

for  $h = 1, 2$ , where  $\lambda_0(u) = f_0(u)/(1 - F_0(u))$  is the hazard function corresponding to  $F_0$ . Thus, conditions (A3)-(A5) impose implicit smoothness assumptions on  $K_h(\cdot)$  and  $F_0(\cdot)$ . These sufficient conditions can be refined in a way that make the required smoothness assumptions more transparent; this will now be done.

In place of (A0)-(A7), we impose the following alternative set of conditions:

- (B1) Conditions (A0) and (A1) hold.
- (B2)  $F_0(\cdot)$  is absolutely continuous, with continuous first and second derivatives; moreover,  $\int_{-\infty}^{\infty} |w|^{2+\delta} f_0(w) dw < \infty$  for some  $\delta > 0$ .
- (B3)  $P\{0 < C_1 \leq C_{max}\} = 1$ , where  $C_{max} < \infty$  (i.e., the censoring variable has finite support) and  $F_0(h(C_{max})) < 1$ . Similarly,  $P\{0 \leq N_1 \leq N_{max}\} = 1$ , where  $N_{max} < \infty$ .
- (B4)  $\text{Var}(W_{1j}^2(\theta_0)) < \infty$  for  $j \geq 1$ .
- (B5) (A4) holds; that is,  $S'(\eta_0)$  is non-singular.

Assumption (B2) implies that  $\lambda_0(\cdot)$  in (A.1) is continuous and differentiable; as a result, both  $K_h(w)$  and  $k_h(w)$ ,  $h = 1, 2$  are continuous and bounded for any  $w$  such that  $F_0(w) < 1$ . Assumption (B3) ensures  $h(C_1 - S_{1,N_1}) < \infty$  and that the summations over  $j$  appearing in (2.14) and (2.15) can involve at most  $N_{max} + 1$  terms (i.e., finite sums). Collectively, (B1)-(B4) imply that  $S(\eta)$  and  $S'(\eta)$  exist for  $\eta \in \mathcal{O}$ ; it now follows by Newey (1991, Corollary 3.1) that (A2) and (A3) hold. Under (B5), we obviously have (A4) and hence consistency by Theorem 2.2.2. As a byproduct, (A7) also holds, since we must have  $S_n(\widehat{\eta}_n) = 0$  as  $n \rightarrow \infty$ .

It can be shown that conditions (B1)-(B4) ensure that both (A5) and (A6) hold. The proof that (A5) holds is straightforward; establishing that (A6) holds is equivalent to showing

$$\sum_{s=1}^p E_{\eta_0} \left[ \{\psi(\eta_0, \mathbb{O}_1)\}_s^2 \right] < \infty, \quad (\text{A.2})$$

where  $\{v\}_s$  denotes the  $s^{th}$  element of a vector  $v$ . Using the notation and results from Sections 2.2.2 and 2.2.3, observe that (A.2) is implied by

$$E \left[ \sum_{s=1}^{p-1} \{E_{\eta_0}[D_{F,1}^*(\eta_0)|\mathbb{O}]\}_s^2 + \left(E_{\eta_0}[D_{F,2}^*(\eta_0)|\mathbb{O}]\right)^2 \right] < \infty,$$

with  $E_{\eta_0}[D_{F,r}^*(\eta_0)|\mathbb{O}]$ ,  $r = 1, 2$  being given in (2.12) and (2.13). Using Jensen's inequality and the fact that  $E_{\eta_0}[D_{F,r}^*(\eta_0)] = 0$ ,  $r = 1, 2$ , this condition is implied by

$$\sum_{s=1}^{p-1} \text{Var}_{\eta_0} [\{D_{F,1}^*(\eta_0)\}_s] + \text{Var}_{\eta_0} [D_{F,2}^*(\eta_0)] < \infty,$$

where  $D_{F,r}^*(\eta_0)$ ,  $r = 1, 2$  are given in (2.7). However, this last condition is guaranteed by conditions (B1)-(B4).

APPENDIX B

LARGE SAMPLE THEORY (CHAPTER 2)

### B.1 Proof of Theorem 2.2.1

We prove the desired result in stages. First, let  $m \geq 1$  be a fixed integer and define

$$D_{F,1,m}(\eta) = \sum_{j=1}^m f_j(\theta)Z_j(\theta) \quad \text{and} \quad D_{F,2,m}(\eta) = \sum_{j=1}^m b_j(\eta)(Z_j^2(\theta) - \sigma^2). \quad (\text{B.1})$$

Let  $\eta_0 = (\theta_0^T, \sigma_0^2)^T$  be the true data generating parameter as defined in Condition (A0). By assumption,  $E[Z_j(\theta_0)|\mathcal{H}_j] = 0$  and  $E[Z_j^2(\theta_0) - \sigma_0^2|\mathcal{H}_j] = 0$ ; moreover,  $b_j(\eta_0)$  and  $f_j(\theta_0)$  are known functions of  $\eta_0$  given  $\mathcal{H}_j$  for each  $j \geq 1$ . Hence, for each fixed  $m > 0$ ,

$$E[D_{F,1,m+1}(\eta_0)|\mathcal{H}_{m+1}] = D_{F,1,m}(\eta_0) + E[f_{m+1}(\theta_0)Z_{m+1}(\theta_0)|\mathcal{H}_{m+1}] = D_{F,1,m}(\eta_0)$$

and, similarly,  $E[D_{F,2,m+1}(\eta_0)|\mathcal{H}_{m+1}] = D_{F,2,m}(\eta_0)$ . It follows that  $\{D_{F,i,k}(\eta_0), k \geq 1\}$ ,  $i = 1, 2$ , form mean zero martingale sequences with respect to  $\{\mathcal{H}_j, j \geq 1\}$ .

The results summarized above imply that (B.1) form a pair of unbiased estimating equations; however, this result is not sufficient to ensure that the estimating functions in (2.7) are unbiased. Towards this end, we next note that the unbiasedness of (B.1) can be immediately generalized to a data-dependent choice of  $m$ . Specifically, let  $\tau > 0$  be a fixed constant, and set

$$D_{F,1}^*(\eta; \tau) = \sum_{j \geq 1} I\{S_{j-1} \leq \tau\} f_j(\theta)Z_j(\theta) \quad (\text{B.2})$$

and

$$D_{F,2}^*(\eta; \tau) = \sum_{j \geq 1} I\{S_{j-1} \leq \tau\} b_j(\eta)(Z_j^2(\theta) - \sigma^2). \quad (\text{B.3})$$

The number of summands in each case is now random, corresponding to the specific choice  $m = N(\tau) + 1$  in (B.1). Using iterated expectation and the facts that  $S_{j-1}$  and  $f_j(\theta_0)$  are known given the information in  $\mathcal{H}_j$ , we see that

$$E[D_{F,1}^*(\eta_0; \tau)] = \sum_{j \geq 1} E \left[ I\{S_{j-1} \leq \tau\} f_j(\theta_0) E[Z_j(\theta_0) | \mathcal{H}_j] \right] = 0.$$

Evidently, this result implies that  $Z_j(\theta_0)$  and  $I\{S_{j-1} \leq \tau\} f_j(\theta_0)$  are conditionally uncorrelated given  $\mathcal{H}_j$ . A similar argument shows  $E[D_{F,2}^*(\eta_0; \tau)] = 0$ .

REMARK: The estimating equations (B.2) and (B.3) depend on the “full” data

$$\mathbb{F}_\tau = \{S_1, S_2, \dots, S_{N(\tau)}, S_{N(\tau)+1}; \bar{L}_1, \dots, \bar{L}_{N(\tau)+1}\},$$

thereby requiring the availability of the observation  $S_{N(\tau)+1}$  (i.e., the time of the first event following time  $\tau$ ). Upon reflection, this requirement is not surprising. Specifically, the random variable  $N(\tau) + 1$ , not  $N(\tau)$ , behaves like a stopping time under the information sequence  $\{\mathcal{H}_j, j \geq 1\}$ . Since martingale behavior is preserved under random stopping (e.g. Fleming and Harrington, 1991, Theorem 2.2.2), one should expect to see that each of (B.2) and (B.3) have mean zero at  $\eta = \eta_0$ .

The unbiasedness of (B.2) and (B.3) is now extended to the setting where the fixed time  $\tau$  is replaced by the random time  $C$ . Specifically, let

$$\mathbb{F} = \{S_1, S_2, \dots, S_N, S_{N+1}; \bar{L}_1, \dots, \bar{L}_{N+1}; C\},$$

which is merely  $\mathbb{F}_\tau$  at  $\tau = C$ , augmented with the additional information on  $C$ . As pointed out in the paper, this structure is also the same as (2.1), augmented with the additional information on the event time  $S_{N+1}$ . Define the new increasing information sequence  $\{H_j, j \geq 1\}$ , where

$$H_j = \mathcal{H}_j \cup \{C \geq S_{j-1}\}.$$

Under the conditional independence assumption specified in Condition (A0), we have  $E[Y_j|H_j] = E[Y_j|\mathcal{H}_j]$  and  $\text{Var}[Y_j|H_j] = \text{Var}[Y_j|\mathcal{H}_j]$  for  $j \geq 1$ . Consequently,  $E[Z_j(\theta_0)|H_j] = 0$  and  $E[Z_j^2(\theta_0)|H_j] = \sigma_0^2$ . Consider now (B.2) and (B.3) with the choice  $\tau = C$ ; that is, the estimating equations specified in (2.7). Using the definition of  $\{H_j, j \geq 1\}$  and proceeding as above, we have

$$E[D_{F,1}^*(\eta_0)] = \sum_{j \geq 1} E[I\{S_{j-1} \leq C\} f_j(\theta_0) E[Z_j(\theta_0)|H_j]] = 0$$

and, similarly,  $E[D_{F,2}^*(\eta_0)] = 0$ . Hence, the  $\mathbb{F}$ -dependent estimating equations (2.7) are unbiased estimators of zero at  $\eta = \eta_0$ , completing the proof.

## B.2 Proof that (2.20) is biased

To understand the difficulty with using (2.20), it is helpful to begin by noting that (2.20) arises as the solution to

$$\frac{1}{n} \sum_{i=1}^n \left\{ \sum_{j=1}^{N_i} (Z_{ij}^2(\widehat{\theta}^{(k+1)}) - \sigma^2) + \left( \frac{\tilde{V}_i^{(k)}}{[V_{i,N_i+1}(\widehat{\theta}^{(k+1)})]^2} - \sigma^2 \right) \right\} = 0.$$

Under the imputation scheme (2.19), this estimating equation is observed to be a special case of  $S_{n,2}^*(\eta) = 0$ , where

$$S_{n,2}^*(\eta) = \frac{1}{n} \sum_{i=1}^n \left\{ \sum_{j=1}^{N_i} b_{ij}(\eta) (W_{ij}^2(\eta) - 1) + b_{i,N_i+1}(\eta) (\text{Var}[W_{i,N_i+1}(\eta)|\mathbb{O}_i] - 1) \right\}. \quad (\text{B.4})$$

However, notice that (2.9) and the development leading up to (2.12) and (2.13) imply that (2.15) is in fact a special case of

$$S_{n,2}^{**}(\eta) = \frac{1}{n} \sum_{i=1}^n \left\{ \sum_{j=1}^{N_i} b_{ij}(\eta) (W_{ij}^2(\eta) - 1) + b_{i,N_i+1}(\eta) (E[W_{i,N_i+1}^2(\eta)|\mathbb{O}_i] - 1) \right\}. \quad (\text{B.5})$$

The estimating equation (B.5) is unbiased. Because

$$E[W_{i,N_i+1}^2(\eta)|\mathbb{O}_i] = \text{Var}[W_{i,N_i+1}(\eta)|\mathbb{O}_i] + (E[W_{i,N_i+1}(\eta)|\mathbb{O}_i])^2,$$

and  $(E[W_{i,N_i+1}(\eta)|\mathbb{O}_i])^2$  typically exceeds zero, it follows that (B.4) is biased. Consequently, the use of  $\widetilde{V}_i^{(k)}$ ,  $i = 1, \dots, n$  in (2.20) leads to a biased estimate  $\sigma^2$ . Furthermore, when the  $N_i$ s are generally small, the terms  $\text{Var}[W_{i,N_i+1}(\eta)|\mathbb{O}_i] - 1$  contribute a larger proportion of the information to the estimating equation, increasing bias. Simulation results (not shown) confirm this intuition.



## APPENDIX C

### GEE: EXTENSIONS TO OTHER WORKING CORRELATION STRUCTURES (CHAPTER 2)

In considering the estimating equations (2.21) and (2.22) as an example of a GEE system, it is interesting to consider the possibility of replacing  $I_{N_i+1}$  appearing in the weight matrix  $A_i(\theta)I_{N_i+1}A_i(\theta)$  with another matrix  $R_i$  that is not necessarily diagonal. For a generic square matrix  $R_i$ , write  $\Omega_i(\eta) = [\sigma A_i(\theta)R_iA_i(\theta)]^{-1}$ . Then,

$$\sum_{i=1}^n G_i(\theta)\Omega_i(\eta)\epsilon_i(\theta) = \sum_{i=1}^n \sum_{k=1}^{N_i+1} Q_{ik}(\eta) (Y_{ik} - \mu_{ik}(\theta)),$$

where  $Q_{ik}(\eta) = \sum_{s=1}^{N_i+1} \frac{d\mu_{is}(\theta)}{d\theta} \Omega_{isk}(\eta)$ . In computing the expectation of the right-hand side, observe that the dependence of each element of the vector  $Q_{ik}(\eta)$  on the full set of gap times  $Y_{ik}$ ,  $k = 1, \dots, N_i + 1$  will in general destroy the unbiasedness of the estimating equation. Importantly, this rules out standard choices of working correlation models (e.g., autoregressive), which involve specifying  $R_i$  as a symmetric matrix. A similar observation can be made for the corresponding generalization of (2.22).

With a bit of reflection, this result is not very surprising: because the moments  $\mu_{ij}(\theta)$  and  $V_{ij}(\theta)$  are defined conditionally on the event history, proper choices of  $R_i$  should reflect this conditional structure. Suppose that the  $r^{th}$  row of  $R_i$  contains the conditional correlations  $\text{Corr}(Y_{ir}, Y_{ik}|\mathcal{H}_{ir})$  for  $k = 1, \dots, N_i + 1$ ,  $r = 1, \dots, N_i + 1$ . Since  $\text{Cov}(Y_{ij}, Y_{i,j+k}|\mathcal{H}_{i,j+k}) = 0$  for  $j, k \geq 1$ , both  $R_i$  and  $\Omega_i(\eta)$  are upper triangular matrices. It follows that  $Q_{ik}(\eta) = \sum_{s=1}^k \frac{d\mu_{is}(\theta)}{d\theta} \Omega_{isk}(\eta)$  for each  $i$  and  $k \geq 1$  and, due to the way in which  $\Omega_i(\eta)$  is constructed,  $Q_{ik}(\eta)$  now depends only on the information available in  $\mathcal{H}_{ik}$ . The resulting estimating equation  $\sum_{i=1}^n G_i(\theta)\Omega_i(\eta)\epsilon_i(\theta)$  and its corresponding projection onto the observed data

are then easily shown to be unbiased.

With the above in hand, it is now evident that the working correlation structure  $R_i = I_{N_i+1}$  underlying (2.21) and (2.22) is not really one of independence but in fact reflects a working assumption that the complete gap times are conditionally uncorrelated. Other correlation structures may be introduced through the use of more general upper triangular matrices  $R_i$ . For example, suppose that  $R_i = T_{N_i+1}(\rho)$ , where  $T_{N_i+1}(\rho)$  is an upper triangular matrix with ones on the diagonal and a constant correlation  $\rho$  in all other non-zero entries. Then,

$$\Omega_i^{-1}(\eta) = \sigma \begin{cases} V_{ik}^2(\theta) & s = k \\ V_{ik}(\theta)V_{is}(\theta)\rho & s < k \\ 0 & s > k \end{cases} \quad \text{and} \quad \Omega_i(\eta) = \sigma^{-1} \begin{cases} \frac{1}{V_{ik}^2(\theta)} & s = k \\ \frac{\rho(-1)^{k-s}(\rho-1)^{k-s-1}}{V_{ik}(\theta)V_{is}(\theta)} & s < k \\ 0 & s > k \end{cases},$$

implying that  $\sum_{i=1}^n G_i(\theta)\Omega_i(\eta)\epsilon_i(\theta) = (I) + (II)$ , where

$$(I) = \frac{1}{\sigma} \sum_{i=1}^n \sum_{k=1}^{N_i+1} \frac{\frac{d\mu_{ik}(\theta)}{d\theta}}{V_{ik}^2(\theta)} (Y_{ik} - \mu_{ik}(\theta))$$

and

$$(II) = \frac{\rho}{\sigma} \sum_{i=1}^n \sum_{k=1}^{N_i+1} \left( \sum_{s=1}^{k-1} \frac{d\mu_{is}(\theta)}{d\theta} \frac{(-1)^{k-s}(\rho-1)^{k-s-1}}{V_{ik}(\theta)V_{is}(\theta)} \right) (Y_{ik} - \mu_{ik}(\theta)).$$

Similar calculations are possible with the corresponding generalization of (2.22); moreover, the working correlation structure here need not be the same as that considered above. The procedure described in Section 2.2.3 for projecting these full data estimating equations onto the observed data remains unchanged.

APPENDIX D

**ADDITIONAL TABLES (CHAPTER 2)**

Table D.1: Simulation results for  $n=50$ ,  $C_i \sim N(225, 50)$

Expected number of events $\doteq 7.4$			True $F_0$					
			Normal			Exponential		
Model	Imputed $F_0$	Parameter	rBias	ESE	ASE	rBias	ESE	ASE
$\mu_{ij}(\theta) = (2.25)$ $V_{ij}(\theta) = (2.24)$	Normal	$\gamma_0$	0.016	0.194	0.192	0.002	0.191	0.191
		$\gamma_1$	0.001	0.292	0.289	0.034	0.292	0.284
		$\rho$	0.057	0.033	0.031	0.156	0.033	0.031
		$\sigma^2$	0.006	0.871	0.836	0.028	1.537	1.492
	Exponential	$\gamma_0$	0.004	0.200	0.191	0.018	0.195	0.191
		$\gamma_1$	0.058	0.288	0.288	0.034	0.280	0.285
		$\rho$	0.093	0.031	0.031	0.087	0.034	0.031
		$\sigma^2$	0.015	0.877	0.880	0.010	1.615	1.544
$\mu_{ij}(\theta) = (2.25)$ $V_{ij}(\theta) =  (2.25) $	Normal	$\gamma_0$	0.006	0.198	0.192	0.011	0.196	0.193
		$\gamma_1$	0.029	0.291	0.289	0.002	0.289	0.284
		$\rho$	0.059	0.034	0.031	0.047	0.036	0.032
		$\sigma^2$	0.008	0.001	0.001	0.034	0.002	0.002
	Exponential	$\gamma_0$	0.004	0.193	0.191	0.009	0.200	0.190
		$\gamma_1$	0.013	0.286	0.289	0.066	0.297	0.284
		$\rho$	0.103	0.033	0.031	0.180	0.033	0.031
		$\sigma^2$	0.012	0.001	0.001	0.015	0.002	0.002

Table D.2: Simulation results for  $n=50$ ,  $C_i \sim N(125, 50)$

Expected number of events $\doteq 3.9$			True $F_0$					
			Normal			Exponential		
Model	Imputed $F_0$	Parameter	rBias	ESE	ASE	rBias	ESE	ASE
$\mu_{ij}(\theta) = (2.25)$ $V_{ij}(\theta) = (2.24)$	Normal	$\gamma_0$	0.036	0.366	0.335	0.022	0.350	0.330
		$\gamma_1$	0.024	0.524	0.496	0.014	0.493	0.473
		$\rho$	0.294	0.052	0.052	0.077	0.059	0.054
		$\sigma^2$	0.007	1.184	1.202	0.050	2.193	1.985
	Exponential	$\gamma_0$	0.008	0.354	0.341	0.021	0.363	0.395
		$\gamma_1$	0.016	0.512	0.500	0.080	0.520	0.544
		$\rho$	0.120	0.057	0.054	0.070	0.057	0.089
		$\sigma^2$	0.030	1.306	1.276	0.003	2.367	2.727
$\mu_{ij}(\theta) = (2.25)$ $V_{ij}(\theta) =  (2.25) $	Normal	$\gamma_0$	0.025	0.346	0.342	0.024	0.380	0.337
		$\gamma_1$	0.084	0.506	0.505	0.037	0.510	0.480
		$\rho$	0.174	0.057	0.054	0.119	0.063	0.055
		$\sigma^2$	0.014	0.001	0.001	0.060	0.002	0.002
	Exponential	$\gamma_0$	0.027	0.361	0.341	0.014	0.345	0.336
		$\gamma_1$	0.052	0.529	0.499	0.017	0.489	0.484
		$\rho$	0.137	0.058	0.054	0.100	0.061	0.056
		$\sigma^2$	0.027	0.002	0.001	0.008	0.003	0.002

Table D.3: Simulation results for  $n=200$ ,  $C_i \sim N(225, 0)$

Expected number of events $\doteq 7.4$			True $F_0$					
Model	Imputed $F_0$	Parameter	Normal			Exponential		
			rBias	ESE	ASE	rBias	ESE	ASE
$\mu_{ij}(\theta) = (2.25)$ $V_{ij}(\theta) = (2.24)$	Normal	$\gamma_0$	0.003	0.096	0.095	0.004	0.101	0.099
		$\gamma_1$	0.000	0.140	0.137	0.013	0.134	0.134
		$\rho$	0.053	0.017	0.016	0.073	0.017	0.017
		$\sigma^2$	0.001	0.442	0.431	0.028	0.828	0.830
	Exponential	$\gamma_0$	0.010	0.091	0.095	0.002	0.096	0.096
		$\gamma_1$	0.002	0.137	0.137	0.007	0.131	0.133
		$\rho$	0.053	0.017	0.017	0.012	0.017	0.017
		$\sigma^2$	0.024	0.456	0.452	0.000	0.849	0.816
$\mu_{ij}(\theta) = (2.25)$ $V_{ij}(\theta) =  (2.25) $	Normal	$\gamma_0$	0.003	0.098	0.096	0.005	0.099	0.099
		$\gamma_1$	0.011	0.137	0.138	0.014	0.135	0.137
		$\rho$	0.022	0.017	0.017	0.069	0.017	0.017
		$\sigma^2$	0.003	0.000	0.001	0.031	0.001	0.001
	Exponential	$\gamma_0$	0.013	0.097	0.095	0.003	0.093	0.096
		$\gamma_1$	0.014	0.137	0.138	0.014	0.132	0.134
		$\rho$	0.061	0.017	0.017	0.006	0.017	0.017
		$\sigma^2$	0.020	0.001	0.001	0.001	0.001	0.001

Table D.4: Simulation results for  $n=200$ ,  $C_i \sim N(225, 50)$

Expected number of events $\doteq 7.4$			True $F_0$					
			Normal			Exponential		
Model	Imputed $F_0$	Parameter	rBias	ESE	ASE	rBias	ESE	ASE
$\mu_{ij}(\theta) = (2.25)$ $V_{ij}(\theta) = (2.24)$	Normal	$\gamma_0$	0.004	0.092	0.097	0.006	0.100	0.104
		$\gamma_1$	0.022	0.147	0.146	0.015	0.143	0.154
		$\rho$	0.013	0.016	0.016	0.068	0.016	0.017
		$\sigma^2$	0.001	0.425	0.428	0.021	0.879	0.949
	Exponential	$\gamma_0$	0.001	0.094	0.097	0.007	0.102	0.097
		$\gamma_1$	0.016	0.147	0.146	0.008	0.150	0.145
		$\rho$	0.015	0.016	0.016	0.033	0.017	0.016
		$\sigma^2$	0.020	0.446	0.444	0.003	0.812	0.814
$\mu_{ij}(\theta) = (2.25)$ $V_{ij}(\theta) =  (2.25) $	Normal	$\gamma_0$	0.004	0.095	0.097	0.008	0.099	0.098
		$\gamma_1$	0.009	0.146	0.147	0.001	0.146	0.144
		$\rho$	0.027	0.017	0.016	0.015	0.017	0.017
		$\sigma^2$	0.000	0.001	0.001	0.028	0.001	0.001
	Exponential	$\gamma_0$	0.008	0.097	0.097	0.007	0.097	0.097
		$\gamma_1$	0.002	0.156	0.146	0.011	0.143	0.145
		$\rho$	0.013	0.017	0.016	0.029	0.017	0.017
		$\sigma^2$	0.020	0.001	0.001	0.003	0.001	0.001

Table D.5: Simulation results for  $n=200$ ,  $C_i \sim N(125, 0)$

Expected number of events $\doteq 3.9$			True $F_0$					
Model	Imputed $F_0$	Parameter	Normal			Exponential		
			rBias	ESE	ASE	rBias	ESE	ASE
$\mu_{ij}(\theta) = (2.25)$ $V_{ij}(\theta) = (2.24)$	Normal	$\gamma_0$	0.034	0.234	0.234	0.032	0.228	0.233
		$\gamma_1$	0.073	0.347	0.355	0.050	0.351	0.352
		$\rho$	0.052	0.031	0.029	0.116	0.033	0.032
		$\sigma^2$	0.001	0.647	0.629	0.030	1.147	1.059
	Exponential	$\gamma_0$	0.008	0.239	0.239	0.005	0.240	0.239
		$\gamma_1$	0.003	0.367	0.356	0.018	0.358	0.358
		$\rho$	0.185	0.030	0.030	0.031	0.032	0.031
		$\sigma^2$	0.017	0.646	0.649	0.002	1.122	1.121
$\mu_{ij}(\theta) = (2.25)$ $V_{ij}(\theta) =  (2.25) $	Normal	$\gamma_0$	0.018	0.227	0.236	0.019	0.236	0.231
		$\gamma_1$	0.047	0.343	0.355	0.023	0.356	0.347
		$\rho$	0.009	0.031	0.030	0.182	0.032	0.031
		$\sigma^2$	0.005	0.001	0.001	0.033	0.001	0.001
	Exponential	$\gamma_0$	0.011	0.235	0.240	0.014	0.236	0.238
		$\gamma_1$	0.050	0.350	0.356	0.019	0.358	0.355
		$\rho$	0.235	0.030	0.030	0.015	0.032	0.032
		$\sigma^2$	0.023	0.001	0.001	0.007	0.001	0.001

Table D.6: Simulation results for  $n=200$ ,  $C_i \sim N(125, 50)$

Expected number of events $\doteq 3.9$			True $F_0$					
Model	Imputed $F_0$	Parameter	Normal			Exponential		
			rBias	ESE	ASE	rBias	ESE	ASE
$\mu_{ij}(\theta) = (2.25)$ $V_{ij}(\theta) = (2.24)$	Normal	$\gamma_0$	0.003	0.173	0.172	0.021	0.170	0.170
		$\gamma_1$	0.008	0.255	0.252	0.051	0.243	0.243
		$\rho$	0.052	0.028	0.027	0.003	0.029	0.028
		$\sigma^2$	0.000	0.636	0.619	0.050	1.047	1.037
	Exponential	$\gamma_0$	0.015	0.172	0.172	0.012	0.173	0.170
		$\gamma_1$	0.040	0.247	0.253	0.035	0.243	0.245
		$\rho$	0.043	0.028	0.028	0.037	0.029	0.028
		$\sigma^2$	0.040	0.693	0.659	0.001	1.177	1.131
$\mu_{ij}(\theta) = (2.25)$ $V_{ij}(\theta) =  (2.25) $	Normal	$\gamma_0$	0.016	0.175	0.173	0.018	0.178	0.173
		$\gamma_1$	0.025	0.251	0.255	0.027	0.249	0.246
		$\rho$	0.065	0.028	0.028	0.017	0.029	0.029
		$\sigma^2$	0.001	0.001	0.001	0.049	0.001	0.001
	Exponential	$\gamma_0$	0.010	0.180	0.174	0.014	0.179	0.171
		$\gamma_1$	0.052	0.249	0.255	0.028	0.253	0.245
		$\rho$	0.037	0.030	0.028	0.000	0.030	0.028
		$\sigma^2$	0.037	0.001	0.001	0.009	0.001	0.001



## APPENDIX E

### VISITATION ASSUMPTIONS (CHAPTER 3)

We will compare the assumptions relating the visitation to the censoring, covariate and outcome processes in five different papers here. Let  $\bar{\mathcal{F}}(t)$  be the observed history of the outcome, covariate, visit, and any auxiliary processes up to and including time  $t$ , let  $\bar{X}(t)$  be the observed history of the outcome process up to and including time  $t$ , let  $A^s(t)$  be a subset of the covariate process  $A$  at time  $t$ , and finally let  $A$  and  $V$  denote a time-fixed covariate and a time-fixed latent variable respectively.

Lipsitz et al. (2002) require  $f(X(t)|\bar{X}(t^-), dN^*(t) = 1) = f(X(t)|\bar{X}(t^-))$ , and by seeing that  $f(dN^*(t)|X(t), \bar{X}(t^-))f(X(t)|\bar{X}(t^-)) = f(dN^*(t)|\bar{X}(t^-))f(X(t)|\bar{X}(t^-), dN^*(t))$  always, this means

$$E(dN^*(t)|X(t), \bar{X}(t^-)) = E(dN^*(t)|\bar{X}(t^-)).$$

They do not consider censoring. Lin et al. (2004), who consider only a time-fixed covariate process (but allow for time-dependent auxiliary variables) require that censoring is independent of the complete auxiliary and outcome processes given the time-fixed covariates. The results for the five papers are summarized in Table E.1. In particular, we can see that the papers after Lin and Ying (2001) in the table allowed the visit process to depend on the histories of processes other than just the covariates from the outcome model. This weaker assumption required the use of inverse-intensity-of-visit-weights for each term

in the estimating equation to make it unbiased. In Lin and Ying (2001),

$$\begin{aligned}
& E \left[ \int_0^\tau (X(t) - E(X(t)|A(t))) dN(t) \right] \\
&= E \left[ \int_0^\tau (X(t) - E(X(t)|A(t))) \xi(t) E(dN^*(t)|A^s(t)) \right] \\
&= E \left[ \int_0^\tau (X(t) - E(X(t)|A(t))) \xi(t) \exp(\gamma^T A^s(t)) d\Lambda_0(t) \right] \\
&= E \left[ \int_0^\tau \xi(t) \exp(\gamma^T A^s(t)) E(X(t) - E(X(t)|A(t))|A(t)) d\Lambda_0(t) \right] \\
&= 0,
\end{aligned}$$

with the last equality following because  $E(X(t) - E(X(t)|A(t))|A(t)) = 0$ . But once  $dN(t)$  is allowed to depend on something other than  $A(t)$ , e.g.  $\{H^{v,m}(t)\}$ , then one is left with

$$E \left[ \int_0^\tau \xi(t) E \left( (X(t) - E(X(t)|A(t))) \exp(\gamma^T h^{v,m}(t)) \right) d\Lambda_0(t) \right] \neq 0,$$

with equality created by adding an inverse-intensity-of-visit-weight to cancel out the  $\exp(\gamma^T h^{v,m}(t))$  term. Equality is also created in our un-weighted conditional estimating equation because of (3.12). See Appendix G.4 for details.

As mentioned in Section 3.2.3, if  $\{H^{v,c}(t)\}$  contains  $X(t)$ , estimation of  $\theta_0^c$  becomes more complicated. Note that in this scenario, (3.12) cannot not hold anymore; therefore  $w^v \neq 1$  is required, much like in (3.5). However, this weight is not enough to remove the bias in the estimate of  $\theta_0^c$  because while this accounts for the informative value of  $X$  at time  $t$ , it does not account for the fact that no observations took place between the previous visit time,  $r(t)$ , and time  $t$ . Equation (3.13) has an implicit term,  $I(N(t^-) - N(r(t)) = 0)$ , which becomes important

Table E.1: A comparison of assumptions regarding the visitation process in five different papers and Chapter 3 of this thesis

Authors	Assumption
Lipsitz et al. (2002)	$E(dN^*(t) X(t), \bar{X}(t^-)) = E(dN^*(t) \bar{X}(t^-))$
Lin and Ying (2001)	$E(dN^*(t) X(t), A(t), C \geq t) = E(dN^*(t) A^s(t))$
Lin et al. (2004)	$E(dN^*(t) X(t), \bar{\mathcal{F}}(t^-)) = E(dN^*(t) \bar{\mathcal{F}}(t^-))$
Sun and Tong (2009)	$E(dN^*(t) X(t), A(t), V, C \geq t) = E(dN^*(t) A, V)$
(3.13)	$E(dN^*(t) X(t), A(t), \bar{\mathcal{F}}(t^-), C \geq t) = E(dN^*(t) \bar{\mathcal{F}}(t^-))$
Bůžková and Lumley (2009)	$E(dN^*(t) X(t), A(t), \bar{\mathcal{F}}(t^-), C \geq t) = E(dN^*(t) A(t), \bar{\mathcal{F}}(t^-))$
(3.5)	$E(dN^*(t) X(t), A(t), \bar{\mathcal{F}}(t^-), C \geq t) = E(dN^*(t) X(t), A(t), \bar{\mathcal{F}}(t^-))$

in this situation. As in Appendix G.4, define

$$BD(t) = \begin{pmatrix} \frac{\partial \mu(t)}{\partial \beta^T} V_1^{-1}(t) & 0 \\ 0 & \frac{\partial \Sigma^*(t)}{\partial \alpha^T} V_2^{-1}(t) \end{pmatrix},$$

$$\epsilon(t) = \begin{pmatrix} X(t) - \mu(t) \\ s^*(t) - \Sigma^*(t) \end{pmatrix},$$

$$I(r(t), t) = I(N(t^-) - N(r(t)) = 0).$$

Then a weighted version of  $U^c(\theta_0^c)$  has limiting expectation

$$\begin{aligned} & E \left[ \int_0^\tau f(h^{m,c}(t)) BD(t) \epsilon(t) \exp(-\gamma_0^T h^{v,c}(t)) dN(t) \right] \\ &= E \left[ \int_0^\tau f(h^{m,c}(t)) BD(t) \epsilon(t) \exp(-\gamma_0^T h^{v,c}(t)) I(r(t), t) dN(t) \right] \\ &= E \left[ \int_0^\tau \xi(t) f(h^{m,c}(t)) BD(t) \epsilon(t) \exp(-\gamma_0^T h^{v,c}(t)) I(r(t), t) \right. \\ &\quad \left. \times E \left( dN^*(t) \middle| H^{v,c}(t), H^{m,c}(t), X(t), \xi(t) = 1 \right) \right] \\ &\stackrel{(3.11)}{=} E \left[ \int_0^\tau \xi(t) f(h^{m,c}(t)) BD(t) \epsilon(t) I(r(t), t) d\Lambda_0(t) \right] \end{aligned}$$

$$\begin{aligned}
&= E \left[ \int_0^\tau \xi(t) f(h^{m,c}(t)) BD(t) \epsilon(t) \right. \\
&\quad \left. \times E \left( I(r(t), t) \middle| H^{m,c}(t), H^{v,c}(t), X(t), A(t), \xi(t) = 1 \right) d\Lambda_0(t) \right] \\
&= E \left[ \int_0^\tau \xi(t) f(h^{m,c}(t)) BD(t) \epsilon(t) \right. \\
&\quad \left. \times P \left( I(r(t), t) = 1 \middle| H^{m,c}(t), H^{v,c}(t), X(t), A(t), \xi(t) = 1 \right) d\Lambda_0(t) \right].
\end{aligned}$$

Clearly another inverse weight is required to get an unbiased estimating equation, and the weight required is the inverse of

$$P(N(t^-) - N(r(t)) = 0 | H^{m,c}(t), H^{v,c}(t), X(t), A(t), \xi(t) = 1); \quad (\text{E.1})$$

that is, the inverse of the probability of a future subject following this information trajectory not visiting between  $r(t)$  and  $t$ . To evaluate (E.1), some joint modeling assumptions on  $N$  and  $X$  must be made, and a parametric assumption on  $X$  might even be required. This would destroy the simplicity and robustness of our method, so we choose not to do this, and instead do not allow the visitation process to depend on the current underlying  $X$ ; see (3.11). Actually it can be argued that this is appropriate in many applications because the decision to test (visit) at a particular time is often made solely based on values of previous measurements. Assumption (3.11) implies that (E.1) is equal to

$$P(N(t^-) - N(r(t)) = 0 | H^{m,c}(t), H^{v,c}(t), \xi(t) = 1).$$

But now by (3.12), the weight is not required in (3.13) to obtain consistency of  $\hat{\theta}^c$ . See the proof of Theorem 3.2.2 in Appendix G.4.

## APPENDIX F

### THEOREMS FOR CONSISTENCY AND ASYMPTOTIC NORMALITY OF SOLUTIONS TO ESTIMATING EQUATIONS (CHAPTERS 3-5)

#### F.1 Notation

Consider  $\hat{\theta}_n$  such that  $\Psi_n(\hat{\theta}_n) = 0$ , for some data-dependent function  $\Psi_n$ ; that is,  $\Psi_n$  is an estimating equation, and  $\hat{\theta}_n$  is a solution to it. Assume that  $\Psi_n$  converges to  $\Psi$  as  $n \rightarrow \infty$ . We assume that  $\Psi$  is a vector-valued function of a parameter  $\theta$ , and that  $\theta$  lies in  $\Theta \subset \mathbb{R}^p$ , for some finite  $p \geq 1$ . It is often desired to show that  $\hat{\theta}_n$  such that  $\Psi_n(\hat{\theta}_n) = 0$  is a consistent and asymptotically normal estimator of the true data-generating parameter  $\theta_0$ . The theorems for consistency and asymptotic normality are given below, and are taken from Kosorok (2008).

#### F.2 Consistency

Let  $\theta_0 \in \Theta$ , and let  $\hat{\theta}_n$  be a sequence of estimators. Assume

$$(F1) \quad \Psi(\theta_0) = 0,$$

$$(F2) \quad \|\Psi(\theta_n)\| \xrightarrow{p} 0 \Rightarrow \|\theta_n - \theta_0\| \xrightarrow{p} 0 \text{ for any deterministic sequence } \theta_n \in \Theta,$$

$$(F3) \quad \|\Psi_n(\hat{\theta}_n)\| \xrightarrow{p} 0,$$

$$(F4) \quad \sup_{\theta \in \Theta} \|\Psi_n(\theta) - \Psi(\theta)\| \xrightarrow{p} 0.$$

Then,  $\hat{\theta}_n$  is a consistent estimator of  $\theta_0$  as  $n \rightarrow \infty$ . Note that (F2) is imposed for identifiability.

### F.3 Asymptotic normality

Let  $\theta_0 \in \Theta$ , and let  $\hat{\theta}_n$  be a sequence of estimators. Assume, as  $n \rightarrow \infty$ ,

(F5)  $n^{1/2}\Psi_n(\hat{\theta}_n) \xrightarrow{p} 0$  and  $\hat{\theta}_n$  is a consistent estimate of  $\theta_0$ ,

(F6)  $n^{1/2}(\Psi_n(\theta_0) - \Psi(\theta_0)) \xrightarrow{d} Z$ , where  $Z \sim \text{MVN}(0, \Xi)$ ,

(F7) 
$$\frac{\|n^{1/2}(\Psi_n(\hat{\theta}_n) - \Psi(\hat{\theta}_n)) - n^{1/2}(\Psi_n(\theta_0) - \Psi(\theta_0))\|}{1 + n^{1/2}\|\hat{\theta}_n - \theta_0\|} \xrightarrow{p} 0,$$

(F8)  $\Psi(\theta)$  is Fréchet differentiable at  $\theta_0$ , i.e. all partials of  $\Psi$  exist and are continuous,

(F9)  $D \equiv \frac{\partial}{\partial \theta} \Psi(\theta) \big|_{\theta_0}$  is non-singular.

Then,

$$n^{1/2}(\hat{\theta}_n - \theta_0) \xrightarrow{d} -D^{-1}Z.$$

APPENDIX G

LARGE SAMPLE THEORY (CHAPTER 3)

**G.1 Regularity conditions sufficient for Theorem 3.2.1**

- (G1) The processes  $X$  and  $A$  are left continuous, and the fourth moments of  $X$  exist.
- (G2) Subjects are independent and identically distributed.
- (G3)  $P(\xi_i(\tau) = 1) > 0$  for all  $i$ .
- (G4)  $N_i(\tau)$  is bounded by a constant for all  $i$ .
- (G5) The hazard of  $N_i^*$  is bounded away from zero for all  $i$  and  $t$ .
- (G6)  $\mu$ ,  $\Sigma$ ,  $V_1^{-1}$ , and  $V_2^{-1}$  are bounded and three times continuously differentiable with respect to  $\theta$  and  $\gamma$ . The third derivatives with respect to  $\theta$  are bounded uniformly in  $\theta$  and the first derivatives with respect to  $\gamma$  are bounded uniformly in  $\gamma$ . Also,  $\mu$  and  $\Sigma$  are correctly specified for all  $t < \tau$ .
- (G7) Let  $\Psi(\theta) = E_{\theta_0^m}(U_1^m(\theta, \gamma_0))$ , where  $U_1^m(\theta, \gamma_0)$  denotes the contribution to (3.5) for one subject. Suppose  $\Psi(\theta)$  satisfies (F1) and (F2) of Appendix F.2. Also let  $\Psi_n(\theta) = n^{-1}U^m(\theta, \hat{\gamma})$ , where  $U^m(\theta, \hat{\gamma})$  is given in (3.5).
- (G8)  $\theta$  and  $\gamma$  are defined on compact sets  $\Theta$  and  $\Gamma$  respectively. The true parameters,  $\theta_0^m$  and  $\gamma_0$ , are assumed to lie interior to  $\Theta$  and  $\Gamma$  respectively.
- (G9) For any sequence  $\gamma_n$  such that  $n^{1/2}(\gamma_n - \gamma_0) = O_p(1)$ , the matrices  $\hat{D}_j$ , for  $j = 2, 3$ , satisfy  $\hat{D}_j = D_j(Id + O_p(n^{-1/2}))$ , where  $Id$  is the identity matrix. Also,  $D_3$  is assumed to be non-singular. These terms are all defined below.
- (G10) The sufficient conditions of Andersen and Gill (1982, Theorem 4.1) hold for the Cox model defined with covariates  $h^{v,m}$ .

(G11) Let  $\Psi(\theta)$  and  $\Psi_n(\theta)$  be defined as in (G7). We assume

$$D_1 = \left. \frac{\partial}{\partial \theta} \Psi(\theta) \right|_{\theta_0^m}$$

exists, is non-singular, and satisfies

$$D_1 \left( \left. \frac{\partial}{\partial \theta} \Psi_n(\theta) \right|_{\hat{\theta}^m} \right)^{-1} \xrightarrow{p} Id.$$

Condition (G5) ensures that an infinite amount of information accumulates for each time interval as  $n \rightarrow \infty$ . Condition (G5) is often stronger than necessary, but if  $\mu$  is a B-spline with interior knots, for example, then information for each inter-knot interval is required for consistency of parameter estimates.

## G.2 Proof of Theorem 3.2.1

In order show consistency of  $\hat{\theta}^m$ , we need to show that the combination of conditions (G1)-(G11) in Appendix G.1 and the conditions outlined directly in Theorem 3.2.1 imply conditions (F1)-(F4) from Appendix F.2 when  $\Psi_n(\theta) \equiv n^{-1}U^m(\theta, \hat{\gamma})$ , as defined in (G7), and the data generating parameter is  $\theta_0^m$ .

Condition (F2) is a direct result of (G7), and (F3) is a direct result of  $\Psi_n(\hat{\theta}^m) \equiv 0$ . Since (G2) assumes subjects are i.i.d.,  $\Psi(\theta_0^m)$  in (G7) has been defined so that we obtain (F1) by showing unbiasedness for just one subject, dropping the associated subject subscripts for convenience. Define

$$BD(t) = \begin{pmatrix} \frac{\partial \mu(t)}{\partial \beta^T} V_1^{-1}(t) & 0 \\ 0 & \frac{\partial \Sigma^*(t)}{\partial \alpha^T} V_2^{-1}(t) \end{pmatrix},$$

$$\epsilon(t) = \begin{pmatrix} X(t) - \mu(t) \\ s^*(t) - \Sigma^*(t) \end{pmatrix}.$$



$$\begin{aligned}
\Psi(\theta_0^m) &= E_{\theta_0^m} \left[ \int_0^\tau f(h^{m,m}(t)) w^v(t, \gamma_0) BD(t) \epsilon(t) dN(t) \right] \\
&= E_{\theta_0^m} \left[ \int_0^\tau E \left( f(h^{m,m}(t)) w^v(t, \gamma_0) BD(t) \epsilon(t) dN(t) \middle| H^{v,m}(t), H^{m,m}(t), X(t), \xi(t) = 1 \right) \right] \\
&= E_{\theta_0^m} \left[ \int_0^\tau \xi(t) f(h^{m,m}(t)) w^v(t, \gamma_0) BD(t) \epsilon(t) \right. \\
&\quad \left. \times E \left( dN^*(t) \middle| H^{v,m}(t), H^{m,m}(t), X(t), \xi(t) = 1 \right) \right] \\
&\stackrel{(3.4)}{=} E_{\theta_0^m} \left[ \int_0^\tau \xi(t) f(h^{m,m}(t)) BD(t) \epsilon(t) d\Lambda_0(t) \right] \\
&= E_{\theta_0^m} \left[ \int_0^\tau \xi(t) f(h^{m,m}(t)) BD(t) E \left( \epsilon(t) \middle| H^{m,m}(t), \xi(t) = 1 \right) d\Lambda_0(t) \right] \\
&\stackrel{(3.3)}{=} E_{\theta_0^m} \left[ \int_0^\tau \xi(t) f(h^{m,m}(t)) BD(t) E \left( \epsilon(t) \middle| H^{m,m}(t) \right) d\Lambda_0(t) \right] \\
&\stackrel{(3.1),(3.2)}{=} 0.
\end{aligned}$$

By the definition in (G7) and a triangle inequality, the left hand side of (F4) in our current setting is bounded by

$$\sup_{\theta \in \Theta} \|\Psi_n(\theta) - n^{-1} U^m(\theta, \gamma_0)\| + \sup_{\theta \in \Theta} \|n^{-1} U^m(\theta, \gamma_0) - \Psi(\theta)\|. \quad (\text{G.1})$$

The first term in (G.1) converges in probability to zero by (G6), (G8), and the consistency of  $\hat{\gamma}$ . The second term also converges in probability to zero by (G1)-(G6) and (G8), which imply the conditions for Nishiyama (2009, Theorem 3.1 (ii)). Therefore (F4) is satisfied and this completes the proof of consistency.

To show asymptotic normality of  $\hat{\theta}^m$  we need to show that the combination of conditions (G1)-(G11) in Appendix G.1 and the conditions outlined directly in Theorem 3.2.1 imply conditions (F5)-(F9) from Appendix F.3 when  $\Psi_n(\theta) \equiv n^{-1} U^m(\theta, \hat{\gamma})$ , as defined in (G7), and the data generating parameter is  $\theta_0^m$ .

Condition (F5) follows directly from the consistency of  $\hat{\theta}^m$  and the fact that  $U^m(\hat{\theta}^m, \hat{\gamma}) \equiv 0$ , and in our current setting, (F8) is a direct consequence of (G6), and

(F9) is a direct consequence of (G11). It remains to show (F6) and (F7), and also to estimate the asymptotic variance of  $\hat{\theta}^m$ . To these ends, a sequence of Taylor series expansions will be helpful. First expand (3.7) around  $\gamma_0$ :

$$0 = U^*(\hat{\gamma}) = U^*(\gamma_0) + \left. \frac{\partial U^*(\gamma)}{\partial \gamma} \right|_{\gamma^*} (\hat{\gamma} - \gamma_0),$$

where  $\gamma^*$  is on the line segment between  $\hat{\gamma}$  and  $\gamma_0$ . This yields

$$\hat{\gamma} - \gamma_0 = \left( - \left. \frac{\partial U^*(\gamma)}{\partial \gamma} \right|_{\gamma^*} \right)^{-1} U^*(\gamma_0).$$

Then expand (3.5) around  $(\theta_0^m, \gamma_0)$ :

$$U^m(\theta_0^m, \hat{\gamma}) = U^m(\theta_0^m, \gamma_0) + \left. \frac{\partial U^m(\theta_0^m, \gamma)}{\partial \gamma} \right|_{\gamma^\circ} (\hat{\gamma} - \gamma_0),$$

where  $\gamma^\circ$  is on the line segment between  $\hat{\gamma}$  and  $\gamma_0$ , which leads to

$$\frac{1}{\sqrt{n}} U^m(\theta_0^m, \hat{\gamma}) = \frac{1}{\sqrt{n}} U^m(\theta_0^m, \gamma_0) - \frac{1}{n} \left. \frac{\partial U^m(\theta_0^m, \gamma)}{\partial \gamma} \right|_{\gamma^\circ} \left( \frac{1}{n} \left. \frac{\partial U^*(\gamma)}{\partial \gamma} \right|_{\gamma^*} \right)^{-1} \frac{1}{\sqrt{n}} U^*(\gamma_0). \quad (\text{G.2})$$

For convenience, define

$$D_2 = \lim_{n \rightarrow \infty} E \left[ - \frac{1}{n} \left. \frac{\partial U^m(\theta_0^m, \gamma)}{\partial \gamma} \right|_{\gamma_0} \right] \quad \text{and} \quad D_3 = \lim_{n \rightarrow \infty} E \left[ - \frac{1}{n} \left. \frac{\partial U^*(\gamma)}{\partial \gamma} \right|_{\gamma_0} \right],$$

and note that by (G9),

$$\hat{D}_2 \equiv - \frac{1}{n} \left. \frac{\partial U^m(\theta_0^m, \gamma)}{\partial \gamma} \right|_{\gamma^\circ} \quad \text{and} \quad \hat{D}_3 \equiv - \frac{1}{n} \left. \frac{\partial U^*(\gamma)}{\partial \gamma} \right|_{\gamma^*}$$

converge at rate  $n^{1/2}$  to  $D_2$  and  $D_3$  respectively.

Some simple algebra shows that

$$U^*(\gamma_0) \equiv \sum_{i=1}^n \int_0^\tau (h_i^{v,m}(t) - \bar{h}_i^{v,m}(t, \gamma_0)) dN_i(t) = \sum_{i=1}^n \int_0^\tau (h_i^{v,m}(t) - \bar{h}_i^{v,m}(t, \gamma_0)) dM_i(t),$$

where  $M_i(t) = N_i(t) - \int_0^t \xi_i(s) \exp(\gamma_0^T h_i^{v,m}(s)) d\Lambda_0(s)$  is a mean zero process. Defining  $U^m(\theta_0^m, \gamma_0) = \sum_{i=1}^n \int_0^\tau \psi_i(t) dN_i(t)$ , we see from (G.2) and (G10) that  $n^{1/2} \Psi_n(\theta_0^m) = n^{-1/2} U^m(\theta_0^m, \hat{\gamma})$  is asymptotically equivalent to

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \int_0^\tau \psi_i(t) dN_i(t) - D_2 D_3^{-1} (h_i^{v,m}(t) - h_0^{v,m}(t, \gamma_0)) dM_i(t), \quad (\text{G.3})$$

where  $h_0^{v,m}(t, \gamma_0)$  is the limit of  $\bar{h}_i^{v,m}(t, \gamma_0)$ . This asymptotic equivalence requires some remainder terms to converge to zero in probability, but this follows from (G9).

It is now evident from (G2) that (G.3) is a sum of i.i.d. mean zero random variables, and so by the central limit theorem, (F6) will be satisfied with

$$Z \sim N\left(0, E\left[\int_0^\tau \psi_1(t) dN_1(t) - D_2 D_3^{-1} (h_1^{v,m}(t) - h_0^{v,m}(t)) dM_1(t)\right]^{\otimes 2}\right),$$

where  $x^{\otimes 2} = xx^T$  for a vector  $x$ , as long as we can show that the summands of (G.3) have finite variance. But the variance of the portion with respect to  $dN_i(t)$  is finite by (G1)'s moment condition and (G4); the variance of the portion with respect to  $dM_i(t)$  is finite by (G1)'s left continuity condition, (G4), and (G10); and all the terms in the covariance of the two portions are finite by the Cauchy-Schwartz inequality and the finite variance results. Note that we also used (F1) to show the second part of the left hand side of (F6) is identically zero.

To show (F7), it helps to note that two of the terms in its numerator are zero. Hence it is sufficient to show:

$$\frac{n^{1/2} \|\Psi(\hat{\theta}^m) - \Psi_n(\theta_0^m)\|}{1 + n^{1/2} \|\hat{\theta}^m - \theta_0^m\|} \xrightarrow{p} 0. \quad (\text{G.4})$$

Some more Taylor expansions are now helpful. First,

$$\Psi(\hat{\theta}^m) = \Psi(\theta_0^m) + D_1(\hat{\theta}^m - \theta_0^m) + \frac{1}{2}(\hat{\theta}^m - \theta_0^m)^T \frac{\partial^2}{\partial \theta \partial \theta^T} \Psi(\theta) \Big|_{\theta^\circ} (\hat{\theta}^m - \theta_0^m), \quad (\text{G.5})$$

where the first term on the right hand side is zero and where  $\theta^\circ$  is on the line segment between  $\hat{\theta}^m$  and  $\theta_0^m$ . Second,

$$0 = U^m(\hat{\theta}^m, \hat{\gamma}) = U^m(\theta_0^m, \hat{\gamma}) + \frac{\partial U^m(\theta, \hat{\gamma})}{\partial \theta} \Big|_{\theta^*} (\hat{\theta}^m - \theta_0^m), \quad (\text{G.6})$$

where  $\theta^*$  is on the line segment between  $\hat{\theta}^m$  and  $\theta_0^m$ . Rearranging (G.6) implies that:

$$\hat{\theta}^m - \theta_0^m = \left( -\frac{\partial U^m(\theta, \hat{\gamma})}{\partial \theta} \Big|_{\theta^*} \right)^{-1} U^m(\theta_0^m, \hat{\gamma}). \quad (\text{G.7})$$

Combining (G.5) and (G.7), and using the triangle inequality, we see that the numerator of the left hand side of (G.4) is bounded by the sum of:

$$\left\| n^{-1/2} \left[ \left( -D_1 \left( -\frac{\partial}{\partial \theta} \Psi_n(\theta) \right) \Big|_{\theta^*} \right)^{-1} - Id \right] U^m(\theta_0^m, \hat{\gamma}) \right\| \quad (G.8)$$

and

$$\left\| \frac{1}{2} n^{1/2} (\hat{\theta}^m - \theta_0^m)^T \frac{\partial^2}{\partial \theta \partial \theta^T} \Psi(\theta) \Big|_{\theta^*} (\hat{\theta}^m - \theta_0^m) \right\|. \quad (G.9)$$

But (G.8) goes in probability to zero by combining (G11) and (F6), and by using the boundedness conditions on the third derivatives in (G6) and the consistency of  $\hat{\theta}^m$ , (G.9) is bounded by a term that is  $o_p(1) n^{1/2} \|\hat{\theta}^m - \theta_0^m\|$ , where  $o_p(1)$  denotes convergence to zero in probability. This demonstrates (G.4), which implies (F7), and so the proof of asymptotic normality is complete.

Finally, using (F6) and the weak convergence result in Appendix F, note that the asymptotic variance of  $n^{1/2}(\hat{\theta}^m - \theta_0^m)$  is:

$$D_1^{-1} \left( E \left[ \int_0^\tau \psi_1(t) dN_1(t) - D_2 D_3^{-1} (h_1^{v,m}(t) - h_0^{v,m}(t)) dM_1(t) \right]^{\otimes 2} \right) (D_1^{-1})^T. \quad (G.10)$$

An estimate of the variance is obtained by replacing the unknown quantities in (G.10) with their corresponding sample quantities.

### G.3 Regularity conditions sufficient for Theorem 3.2.2

(H1) Conditions (G1)-(G5) in Appendix G.1 hold.

(H2)  $\mu$ ,  $\Sigma$ ,  $V_1^{-1}$ , and  $V_2^{-1}$  are bounded and three times continuously differentiable with respect to  $\theta$ . The third derivatives are bounded uniformly in  $\theta$ . Also,  $\mu$  and  $\Sigma$  are correctly specified for all  $t < \tau$ .

(H3) Let  $\Psi(\theta) = E_{\theta_0^c}(U_1^c(\theta))$ , where  $U_1^c(\theta)$  denotes the contribution to (3.13) for one subject. Suppose  $\Psi(\theta)$  satisfies (F1) and (F2) of Appendix F.2. Also let  $\Psi_n(\theta) = n^{-1}U^c(\theta)$ , where  $U^c(\theta)$  is given in (3.13).

(H4)  $\theta$  is defined on a compact set  $\Theta$ . The true parameter,  $\theta_0^c$  is assumed to lie interior to  $\Theta$ .

(H5) Let  $\Psi(\theta)$  and  $\Psi_n(\theta)$  be defined as in (H3). We assume

$$D_4 = \left. \frac{\partial}{\partial \theta} \Psi(\theta) \right|_{\theta_0^c}$$

exists, is non-singular, and satisfies

$$D_4 \left( \left. \frac{\partial}{\partial \theta} \Psi_n(\theta) \right|_{\hat{\theta}^c} \right)^{-1} \xrightarrow{p} Id,$$

where  $Id$  is the identity matrix.

## G.4 Proof of Theorem 3.2.2

In order show consistency of  $\hat{\theta}^c$ , we need to show that the combination of conditions (H1)-(H5) in Appendix G.3 and the conditions outlined directly in Theorem 3.2.2 imply conditions (F1)-(F4) from Appendix F.2 when  $\Psi_n(\theta) \equiv n^{-1}U^c(\theta)$ , as defined in (H3), and the data generating parameter is  $\theta_0^c$ .

Condition (F2) is a direct result of (H3), (F3) is a direct result of  $\Psi_n(\hat{\theta}^c) \equiv 0$ , and (F4) holds because (H1), (H2), and (H4) imply the conditions for Nishiyama (2009, Theorem 3.1 (ii)). It only remains to show (F1).

Since (G2) assumes subjects are i.i.d.,  $\Psi(\theta_0^c)$  in (H3) has been defined so that we obtain (F1) by showing unbiasedness for just one subject, dropping the associated subject subscripts for convenience. Define  $BD(t)$  and  $\epsilon(t)$  as the CEE

analogous of those terms from Appendix G.2. Also define  $r(t) = \max(s : dN(s) = 1, s < t)$  and  $I(r(t), t) = I(N(t^-) - N(r(t)) = 0)$ , and let  $h^{v,c}(t)$  be the covariates in the hazard for visitation at time  $t$ .

$$\begin{aligned}
\Psi(\theta_0^c) &= E_{\theta_0^c} \left[ \int_0^\tau f(h^{m,c}(t)) BD(t) \epsilon(t) dN(t) \right] \\
&= E_{\theta_0^c} \left[ \int_0^\tau f(h^{m,c}(t)) BD(t) \epsilon(t) I(r(t), t) dN(t) \right] \\
&= E_{\theta_0^c} \left[ \int_0^\tau E \left( f(h^{m,c}(t)) BD(t) \epsilon(t) I(r(t), t) dN(t) \middle| H^{v,c}(t), H^{m,c}(t), X(t), \xi(t) = 1 \right) \right] \\
&= E_{\theta_0^c} \left[ \int_0^\tau \xi(t) f(h^{m,c}(t)) BD(t) \epsilon(t) I(r(t), t) \right. \\
&\quad \left. \times E \left( dN^*(t) \middle| H^{v,c}(t), H^{m,c}(t), X(t), \xi(t) = 1 \right) \right] \\
&\stackrel{(3.11)}{=} E_{\theta_0^c} \left[ \int_0^\tau \xi(t) f(h^{m,c}(t)) BD(t) \epsilon(t) I(r(t), t) \exp(\gamma^T h^{v,c}(t)) d\Lambda_0(t) \right] \\
&= E_{\theta_0^c} \left[ \int_0^\tau \xi(t) f(h^{m,c}(t)) BD(t) \exp(\gamma^T h^{v,c}(t)) I(r(t), t) \right. \\
&\quad \left. \times E \left( \epsilon(t) \middle| H^{m,c}(t), H^{v,c}(t), \xi(t) = 1, I(r(t), t) \right) d\Lambda_0(t) \right] \\
&\stackrel{(3.11)}{=} E_{\theta_0^c} \left[ \int_0^\tau \xi(t) f(h^{m,c}(t)) BD(t) \exp(\gamma^T h^{v,c}(t)) I(r(t), t) \right. \\
&\quad \left. \times E \left( \epsilon(t) \middle| H^{m,c}(t), H^{v,c}(t), \xi(t) = 1 \right) d\Lambda_0(t) \right] \\
&\stackrel{(3.12)}{=} E_{\theta_0^c} \left[ \int_0^\tau \xi(t) f(h^{m,c}(t)) BD(t) \exp(\gamma^T h^{v,c}(t)) I(r(t), t) \right. \\
&\quad \left. \times E \left( \epsilon(t) \middle| H^{m,c}(t), \xi(t) = 1 \right) d\Lambda_0(t) \right] \\
&\stackrel{(3.10)}{=} E_{\theta_0^c} \left[ \int_0^\tau \xi(t) f(h^{m,c}(t)) BD(t) \exp(\gamma^T h^{v,c}(t)) I(r(t), t) E \left( \epsilon(t) \middle| H^{m,c}(t) \right) d\Lambda_0(t) \right] \\
&\stackrel{(3.8), (3.9)}{=} 0.
\end{aligned}$$

This completes the proof of consistency.

To show asymptotic normality of  $\hat{\theta}^c$  we need to show that the combination of conditions (H1)-(H5) in Appendix G.3 and the conditions outlined directly in Theorem 3.2.2 imply conditions (F5)-(F9) from Appendix F.3 when  $\Psi_n(\theta) \equiv$

$n^{-1}U^c(\theta)$ , as defined in (H3), and the data generating parameter is  $\theta_0^c$ . Condition (F5) follows directly from the consistency of  $\hat{\theta}^c$  and the fact that  $U^c(\hat{\theta}^c) \equiv 0$ , and in our current setting, (F8) is a direct consequence of (H2), and (F9) is a direct consequence of (H5). It remains to show (F6) and (F7), and also to estimate the asymptotic variance of  $\hat{\theta}^c$ .

It is evident from (G2) that  $U^c(\theta_0^c)$  is a sum of i.i.d. mean zero random variables. Defining  $U^c(\theta_0^c) = \sum_{i=1}^n \int_0^\tau \psi_i(t) dN_i(t)$ , we see from the central limit theorem that (F6) will be satisfied with

$$Z \sim N\left(0, E\left[\int_0^\tau \psi_1(t) dN_1(t)\right]^{\otimes 2}\right),$$

where  $x^{\otimes 2} = xx^T$  for a vector  $x$ , as long as we can show that the summands of  $U^c(\theta_0^c)$  have finite variance. But this follows from (G1)'s moment condition and (G4). Note that we also used (F1) to show the second part of the left hand side of (F6) is identically zero.

To show (F7), it helps to note that two of the terms in its numerator are zero. Hence it is sufficient to show:

$$\frac{n^{1/2}\|\Psi(\hat{\theta}^c) - \Psi_n(\theta_0^c)\|}{1 + n^{1/2}\|\hat{\theta}^c - \theta_0^c\|} \xrightarrow{p} 0. \quad (\text{G.11})$$

Some Taylor expansions are now helpful. First,

$$\Psi(\hat{\theta}^c) = \Psi(\theta_0^c) + D_4(\hat{\theta}^c - \theta_0^c) + \frac{1}{2}(\hat{\theta}^c - \theta_0^c)^T \frac{\partial^2}{\partial\theta\partial\theta^T} \Psi(\theta)|_{\theta^\circ} (\hat{\theta}^c - \theta_0^c), \quad (\text{G.12})$$

where the first term on the right hand side is zero and where  $\theta^\circ$  is on the line segment between  $\hat{\theta}^c$  and  $\theta_0^c$ . Second,

$$0 = U^c(\hat{\theta}^c) = U^c(\theta_0^c) + \frac{\partial U^c(\theta)}{\partial\theta} \Big|_{\theta^*} (\hat{\theta}^c - \theta_0^c), \quad (\text{G.13})$$

where  $\theta^*$  is on the line segment between  $\hat{\theta}^c$  and  $\theta_0^c$ . Rearranging (G.13) implies that:

$$\hat{\theta}^c - \theta_0^c = \left(-\frac{\partial U^c(\theta)}{\partial\theta} \Big|_{\theta^*}\right)^{-1} U^c(\theta_0^c). \quad (\text{G.14})$$

Combining (G.12) and (G.14), and using the triangle inequality, we see that the numerator of the left hand side of (G.11) is bounded by the sum of:

$$\left\| n^{-1/2} \left[ \left( -D_4 \left( -\frac{\partial}{\partial \theta} \Psi_n(\theta) \Big|_{\theta^c} \right)^{-1} - Id \right) U^c(\theta_0^c) \right] \right\| \quad (G.15)$$

and

$$\left\| \frac{1}{2} n^{1/2} (\hat{\theta}^c - \theta_0^c)^T \frac{\partial^2}{\partial \theta \partial \theta^T} \Psi(\theta) \Big|_{\theta_0^c} (\hat{\theta}^c - \theta_0^c) \right\|. \quad (G.16)$$

But (G.15) goes in probability to zero by combining (H5) and (F6), and by using the boundedness conditions on the third derivatives in (H2) and the consistency of  $\hat{\theta}^c$ , (G.16) is bounded by a term that is  $o_p(1) n^{1/2} \|\hat{\theta}^c - \theta_0^c\|$ , where  $o_p(1)$  denotes convergence to zero in probability. This demonstrates (G.11), which implies (F7), and so the proof of asymptotic normality is complete.

Finally, using (F6) and the weak convergence result in Appendix F, note that the asymptotic variance of  $n^{1/2}(\hat{\theta}^c - \theta_0^c)$  is:

$$D_4^{-1} \left( E \left[ \int_0^\tau \psi_1(t) dN_1(t) \right]^{\otimes 2} \right) (D_4^{-1})^T. \quad (G.17)$$

An estimate of the variance is obtained by replacing the unknown quantities in (G.17) with their corresponding sample quantities.



## APPENDIX H

### MISSING DATA TRANSITION DENSITY (CHAPTER 4)

We will show a derivation of the conditional transition density of a multivariate Ornstein-Uhlenbeck process here. Conditioning on survival to the appropriate time is necessary, but has been left out of the notation for simplicity. Recall that for the Ornstein-Uhlenbeck process:

$$dX(t) = B(\mu - X(t))dt + \Sigma^{1/2}dW(t),$$

and

$$(X(s+t)|X(s), s) \sim N\left(\mu + \exp(-Bt)(X(s) - \mu), \Sigma^\dagger - \exp(-Bt)\Sigma^\dagger \exp(-B^T t)\right),$$

where  $\mu = (\mu_1, \dots, \mu_d)$  is an unknown parameter vector and  $B$  and  $\Sigma^\dagger = (2B)^{-1/2}\Sigma(2B)^{-1/2} = [(\sigma)_{ij}]$ ,  $i, j = 1, \dots, d$  are unknown non-negative definite matrices of parameters. The exponential of a matrix is defined as the power series,  $\exp(B) = \sum_{k=0}^{\infty} \frac{1}{k!} B^k$ , which is hard to compute in practice. Fortunately if  $B$  is a  $d \times d$  diagonal matrix,  $\exp(B)$  is diagonal with entries  $\exp(b_1), \dots, \exp(b_d)$ . The higher  $b_j$  is, the faster the  $j^{\text{th}}$  covariate reverts to its mean,  $\mu_j$ . We will work with a diagonal  $B$  for computational convenience. The formulas that follow would be possible to extend to a non-diagonal  $B$ , but would be more cumbersome, and would have to be approximated in order to do (the intensive) computation.

It desired to know the distribution of  $X(t)$  conditioned on  $X(r^*(t)) \equiv (X_1(r_1^*(t)), \dots, X_d(r_d^*(t)))$ . But the diagonal form of  $B$  and the independent increments make the calculation easy: the multivariate mean is just the vector of the marginal means of each dimension and the covariance between dimensions  $i$  and  $j$  is the same as the covariance over the interval  $(\max(r_i^*(t), r_j^*(t)), t)$ . This

gives:

$$(X(t)|X(r^*(t)) \sim N\left(\mu + \exp(-B(\tilde{t} - r^*(t)))(X(r^*(t)) - \mu), \tilde{\Sigma}_{ij}(t)\right),$$

where  $\tilde{t}$  denotes a  $d \times 1$  vector with  $t$  repeated  $d$  times and  $\tilde{\Sigma}_{ij}(t) = \sigma_{ij}(1 - \exp(-(b_i + b_j)(t - \max(r_i^*(t), r_j^*(t)))))$ . Note that a subject's time-fixed covariates can easily be combined with  $\mu$  or  $B$  to allow different means or different rates of mean reversion between subjects. The deterministic drift model derivation is similar.

## APPENDIX I

### CONDITIONAL EXPECTATION CALCULATIONS (CHAPTER 4)

There are four different cases for the expected value calculations in (4.9), all relying on (4.5). The first relates to the mean estimating equation, and is simply

$$E\left(X_i^u(t)|X_i^\dagger(t)\right) = \tilde{\mu}_{i1}(t) + \tilde{\Sigma}_{i12}(t)\tilde{\Sigma}_{i22}^{-1}(t)(c - \tilde{\mu}_{i2}(t)) \equiv \tilde{\mu}_{i1}\left(t|X_i^\dagger(t)\right)$$

The remaining three relate to the variance estimating equation.

$$\begin{aligned} & E\left(s_i(t)|X_i^\dagger(t)\right) \\ &= E\left[\begin{pmatrix} (X_i^u(t) - \tilde{\mu}_{i1}(t))(X_i^u(t) - \tilde{\mu}_{i1}(t))^T & (X_i^u(t) - \tilde{\mu}_{i1}(t))(X_i^o(t) - \tilde{\mu}_{i2}(t))^T \\ (X_i^o(t) - \tilde{\mu}_{i2}(t))(X_i^u(t) - \tilde{\mu}_{i1}(t))^T & (X_i^o(t) - \tilde{\mu}_{i2}(t))(X_i^o(t) - \tilde{\mu}_{i2}(t))^T \end{pmatrix} \middle| X_i^\dagger(t)\right] \quad (\text{I.1}) \end{aligned}$$

The bottom right block is just complete data and the off diagonal blocks are the same, so only the equations for the top row of (I.1) are needed.

$$\begin{aligned} & E\left((X_i^u(t) - \tilde{\mu}_{i1}(t))(X_i^u(t) - \tilde{\mu}_{i1}(t))^T | X_i^\dagger(t)\right) \\ &= \text{Cov}\left(X_i^u(t) - \tilde{\mu}_{i1}(t), X_i^u(t) - \tilde{\mu}_{i1}(t) | X_i^\dagger(t)\right) \\ &\quad + E\left(X_i^u(t) - \tilde{\mu}_{i1}(t) | X_i^\dagger(t)\right) E\left(X_i^u(t) - \tilde{\mu}_{i1}(t) | X_i^\dagger(t)\right)^T \\ &= \tilde{\Sigma}_{i11}(t) - \tilde{\Sigma}_{i12}(t)\tilde{\Sigma}_{i22}^{-1}(t)\tilde{\Sigma}_{i21}(t) + \left(\tilde{\mu}_{i1}(t|X_i^\dagger(t)) - \tilde{\mu}_{i1}(t)\right)\left(\tilde{\mu}_{i1}(t|X_i^\dagger(t)) - \tilde{\mu}_{i1}(t)\right)^T \end{aligned}$$

and

$$\begin{aligned} & E\left((X_i^u(t) - \tilde{\mu}_{i1}(t))(X_i^o(t) - \tilde{\mu}_{i2}(t))^T | X_i^\dagger(t)\right) \\ &= E\left(X_i^u(t) - \tilde{\mu}_{i1}(t) | X_i^\dagger(t)\right) (X_i^o(t) - \tilde{\mu}_{i2}(t))^T \\ &= \left(\tilde{\mu}_{i1}(t|X_i^\dagger(t)) - \tilde{\mu}_{i1}(t)\right) (X_i^o(t) - \tilde{\mu}_{i2}(t))^T \end{aligned}$$

APPENDIX J

LARGE SAMPLE THEORY (CHAPTER 4)

**J.1 Regularity conditions sufficient for Theorem 4.2.1**

- (J1) Conditions (H1) and (H2) in Appendix G.3 hold.
- (J2) Regarding missing data, given a visit at time  $t$ , each pair of dimensions of the response process must have a probability bounded away from zero of being observed simultaneously; this must hold uniformly for all  $t$ .
- (J3) Let  $\Psi(\theta) = E_{\theta_0}(\tilde{U}_1^c(\theta))$ , where  $\tilde{U}_1^c(\theta)$  denotes the contribution to (4.9) for one subject. Suppose  $\Psi(\theta)$  satisfies (F1) and (F2) of Appendix F.2. Also let  $\Psi_n(\theta) = n^{-1} \tilde{U}^c(\theta)$ , where  $\tilde{U}^c(\theta)$  is given in (4.9).
- (J4)  $\theta$  is defined on a compact set  $\Theta$ . The true parameter,  $\theta_0$  is assumed to lie interior to  $\Theta$ .
- (J5) Let  $\Psi(\theta)$  and  $\Psi_n(\theta)$  be defined as in (J3). We assume

$$D_5 = \left. \frac{\partial}{\partial \theta} \Psi(\theta) \right|_{\theta_0}$$

exists, is non-singular, and satisfies

$$D_5 \left( \left. \frac{\partial}{\partial \theta} \Psi_n(\theta) \right|_{\hat{\theta}} \right)^{-1} \xrightarrow{p} Id,$$

where  $Id$  is the identity matrix.

Condition (J2) ensures that an infinite amount of information to estimate all dimensions of  $\theta_0$  accumulates as  $n \rightarrow \infty$ .

## J.2 Proof of Theorem 4.2.1

In order to show consistency of  $\hat{\theta}$ , we need to show that the combination of conditions (J1)-(J5) in Appendix J.1 and the conditions outlined directly in Theorem 4.2.1 imply conditions (F1)-(F4) from Appendix F.2 when  $\Psi_n(\theta) \equiv n^{-1} \tilde{U}^c(\theta)$ , as defined in (J3), and the data generating parameter is  $\theta_0$ .

Condition (F2) is a direct result of (J3), (F3) is a direct result of  $\Psi_n(\hat{\theta}) \equiv 0$ , and (F4) holds because (J1) and (J4) imply the conditions for Nishiyama (2009, Theorem 3.1 (ii)). It only remains to show (F1).

Since (G2) assumes subjects are i.i.d.,  $\Psi(\theta_0)$  in (J3) has been defined so that we obtain (F1) by showing unbiasedness for just one subject, dropping the associated subject subscripts for convenience. Define

$$\begin{aligned} \tilde{B}D(t) &\equiv \begin{pmatrix} \frac{\partial \tilde{\mu}(t)}{\partial \beta^T} \tilde{V}_1^{-1}(t) & 0 \\ 0 & \frac{\partial \tilde{\Sigma}^*(t)}{\partial \alpha^T} \tilde{V}_2^{-1}(t) \end{pmatrix}, \\ \epsilon(t) &= \begin{pmatrix} E(X(t)|X^\dagger(t)) - \tilde{\mu}(t) \\ E(\tilde{S}^*(t)|X^\dagger(t)) - \tilde{\Sigma}^*(t) \end{pmatrix}, \end{aligned}$$

$$I(r(t), t) = I(N(t^-) - N(r(t)) = 0),$$

where  $r(t) = \max(s : dN(s) = 1, s < t)$ . Also let  $h^{v,c}(t)$  be the covariates in the hazard for visitation.

$$\begin{aligned} \Psi(\theta_0) &= E_{\theta_0} \left[ \int_0^\tau f(h^{m,c}(t)) \tilde{B}D(t) \epsilon(t) dN(t) \right] \\ &= E_{\theta_0} \left[ \int_0^\tau f(h^{m,c}(t)) \tilde{B}D(t) \epsilon(t) I(r(t), t) dN(t) \right] \\ &= E_{\theta_0} \left[ \int_0^\tau E \left( f(h^{m,c}(t)) \tilde{B}D(t) \epsilon(t) I(r(t), t) dN(t) \middle| H^{v,c}(t), H^{m,c}(t), X^o(t), \xi(t) = 1 \right) \right] \end{aligned}$$

$$\begin{aligned}
&= E_{\theta_0} \left[ \int_0^\tau \xi(t) f(h^{m,c}(t)) \tilde{B}D(t) \epsilon(t) I(r(t), t) \right. \\
&\quad \left. \times E \left( dN^*(t) \middle| H^{v,c}(t), H^{m,c}(t), X^o(t), \xi(t) = 1 \right) \right] \\
&\stackrel{(4.7)}{=} E_{\theta_0} \left[ \int_0^\tau \xi(t) f(h^{m,c}(t)) \tilde{B}D(t) \epsilon(t) I(r(t), t) \exp(\gamma^T h^{v,c}(t)) d\Lambda_0(t) \right] \\
&= E_{\theta_0} \left[ \int_0^\tau \xi(t) f(h^{m,c}(t)) \tilde{B}D(t) \exp(\gamma^T h^{v,c}(t)) I(r(t), t) \right. \\
&\quad \left. \times E \left( \epsilon(t) \middle| H^{m,c}(t), H^{v,c}(t), \xi(t) = 1, I(r(t), t) \right) d\Lambda_0(t) \right] \\
&\stackrel{(4.7)}{=} E_{\theta_0} \left[ \int_0^\tau \xi(t) f(h^{m,c}(t)) \tilde{B}D(t) \exp(\gamma^T h^{v,c}(t)) I(r(t), t) \right. \\
&\quad \left. \times E \left( \epsilon(t) \middle| H^{m,c}(t), H^{v,c}(t), \xi(t) = 1 \right) d\Lambda_0(t) \right] \\
&\stackrel{(4.8)}{=} E_{\theta_0} \left[ \int_0^\tau \xi(t) f(h^{m,c}(t)) \tilde{B}D(t) \exp(\gamma^T h^{v,c}(t)) I(r(t), t) \right. \\
&\quad \left. \times E \left( \epsilon(t) \middle| H^{m,c}(t), \xi(t) = 1 \right) d\Lambda_0(t) \right] \\
&\stackrel{(4.6)}{=} E_{\theta_0} \left[ \int_0^\tau \xi(t) f(h^{m,c}(t)) \tilde{B}D(t) \exp(\gamma^T h^{v,c}(t)) I(r(t), t) E \left( \epsilon(t) \middle| H^{m,c}(t), T \geq t \right) d\Lambda_0(t) \right] \\
&\stackrel{(4.1), (4.2)}{=} 0.
\end{aligned}$$

This completes the proof of consistency.

To show asymptotic normality of  $\hat{\theta}$  we need to show that the combination of conditions (J1)-(J5) in Appendix J.1 and the conditions outlined directly in Theorem 4.2.1 imply conditions (F5)-(F9) from Appendix F.3 when  $\Psi_n(\theta) \equiv n^{-1} \tilde{U}^c(\theta)$ , as defined in (J3), and the data generating parameter is  $\theta_0$ . Condition (F5) follows directly from the consistency of  $\hat{\theta}$  and the fact that  $\tilde{U}^c(\hat{\theta}) \equiv 0$ , and in our current setting, (F8) is a direct consequence of (H2), and (F9) is a direct consequence of (J5). It remains to show (F6) and (F7), and also to estimate the asymptotic variance of  $\hat{\theta}$ .

It is evident from (G2) that  $\tilde{U}^c(\theta_0)$  is a sum of i.i.d. mean zero random variables. Defining  $\tilde{U}^c(\theta_0) = \sum_{i=1}^n \int_0^\tau \psi_i(t) dN_i(t)$ , we see from the central limit theorem

that (F6) will be satisfied with

$$Z \sim N\left(0, E\left[\int_0^\tau \psi_1(t) dN_1(t)\right]^{\otimes 2}\right),$$

where  $x^{\otimes 2} = xx^T$  for a vector  $x$ , as long as we can show that the summands of  $\tilde{U}^c(\theta_0)$  have finite variance. But this follows from (G1)'s moment condition and (G4). Note that we also used (F1) to show the second part of the left hand side of (F6) is identically zero.

To show (F7), it helps to note that two of the terms in its numerator are zero. Hence it is sufficient to show:

$$\frac{n^{1/2}\|\Psi(\hat{\theta}) - \Psi_n(\theta_0)\|}{1 + n^{1/2}\|\hat{\theta} - \theta_0\|} \xrightarrow{p} 0. \quad (\text{J.1})$$

Some Taylor expansions are now helpful. First,

$$\Psi(\hat{\theta}) = \Psi(\theta_0) + D_5(\hat{\theta} - \theta_0) + \frac{1}{2}(\hat{\theta} - \theta_0)^T \frac{\partial^2}{\partial\theta\partial\theta^T} \Psi(\theta)|_{\theta^\circ} (\hat{\theta} - \theta_0), \quad (\text{J.2})$$

where the first term on the right hand side is zero and where  $\theta^\circ$  is on the line segment between  $\hat{\theta}$  and  $\theta_0$ . Second,

$$0 = \tilde{U}^c(\hat{\theta}) = \tilde{U}^c(\theta_0) + \frac{\partial \tilde{U}^c(\theta)}{\partial\theta} \Big|_{\theta^*} (\hat{\theta} - \theta_0), \quad (\text{J.3})$$

where  $\theta^*$  is on the line segment between  $\hat{\theta}$  and  $\theta_0$ . Rearranging (J.3) implies that:

$$\hat{\theta} - \theta_0 = \left(-\frac{\partial \tilde{U}^c(\theta)}{\partial\theta} \Big|_{\theta^*}\right)^{-1} \tilde{U}^c(\theta_0). \quad (\text{J.4})$$

Combining (J.2) and (J.4), and using the triangle inequality, we see that the numerator of the left hand side of (J.1) is bounded by the sum of:

$$\left\| n^{-1/2} \left[ \left( -D_5 \left( -\frac{\partial}{\partial\theta} \Psi_n(\theta) \Big|_{\theta^*} \right)^{-1} - Id \right) \tilde{U}^c(\theta_0) \right] \right\| \quad (\text{J.5})$$

and

$$\left\| \frac{1}{2} n^{1/2} (\hat{\theta} - \theta_0)^T \frac{\partial^2}{\partial\theta\partial\theta^T} \Psi(\theta)|_{\theta^\circ} (\hat{\theta} - \theta_0) \right\|. \quad (\text{J.6})$$

But (J.5) goes in probability to zero by combining (J5) and (F6), and by using the boundedness conditions on the third derivatives in (H2) and the consistency of  $\hat{\theta}$ , (J.6) is bounded by a term that is  $o_P(1)n^{1/2}\|\hat{\theta} - \theta_0\|$ , where  $o_P(1)$  denotes convergence to zero in probability. This demonstrates (J.1), which implies (F7), and so the proof of asymptotic normality is complete.

Finally, using (F6) and the weak convergence result in Appendix F, note that the asymptotic variance of  $n^{1/2}(\hat{\theta} - \theta_0)$  is:

$$D_5^{-1} \left( E \left[ \int_0^\tau \psi_1(t) dN_1(t) \right]^{\otimes 2} \right) (D_5^{-1})^T. \quad (\text{J.7})$$

An estimate of the variance is obtained by replacing the unknown quantities in (J.7) with their corresponding sample quantities.



APPENDIX K  
LARGE SAMPLE THEORY (CHAPTER 5)

**K.1 Regularity conditions sufficient for Theorem 5.3.1**

(K1) Let  $\Psi_n(\kappa) = n^{-1}U^\dagger(\kappa, \hat{\theta})$  and  $\Psi_n^*(\kappa) = n^{-1}U^\dagger(\kappa, \theta_0)$  where  $U^\dagger(\kappa, \hat{\theta})$  is defined in (5.4). Assume that  $\Psi_n(\kappa)$  converges uniformly to a limiting function  $\Psi(\kappa)$ , which is defined below in equation (K.4). We assume  $\Psi(\kappa)$  satisfies (F1) and (F2) of Appendix F.2.

(K2)  $\kappa$  is defined on a compact set  $\Pi$ . The true parameter,  $\kappa_0$  is assumed to lie interior to  $\Pi$ .

(K3) The sufficient conditions of Andersen and Gill (1982, Theorem 4.1) hold for the Cox model defined with covariates  $h^d$ . These conditions are augmented with:

$$\sup_{t < \tau, \theta \in \Theta, \kappa \in \Pi} \|S^{(j)}(t, \kappa, \theta) - s^{(j)}(t, \kappa, \theta)\| \xrightarrow{p} 0,$$

for  $j = 0, 1$ , where  $s^{(1)}$  and  $s^{(0)}$  are vector and scalar functions respectively, and where  $S^{(1)}$  and  $S^{(0)}$  are defined in (K.1) and (K.2) respectively. We also assume that  $U^\dagger(\kappa, \theta)$  is twice continuously differentiable with respect to  $\theta$  and  $\kappa$ , and that the second derivatives with respect to  $\kappa$  are bounded uniformly in  $\kappa$ .

(K4) For any sequence  $\theta_n$  such that  $n^{1/2}(\theta_n - \theta_0) = O_p(1)$ , the matrices  $\hat{D}_j$ , for  $j = 5, 7$ , satisfy  $\hat{D}_j = D_j(Id + O_p(n^{-1/2}))$ , where  $Id$  is the identity matrix. Also,  $D_5$  is assumed to be non-singular. These terms are all defined below.

(K5) Let  $\Psi_n(\kappa)$  and  $\Psi(\kappa)$  be defined as in (K1) and equation (K.4) respectively.

We assume

$$D_6 = \left. \frac{\partial}{\partial \kappa} \Psi(\kappa) \right|_{\kappa_0}$$

is non-singular and satisfies

$$D_6 \left( \left. \frac{\partial}{\partial \kappa} \Psi_n(\kappa) \right|_{\hat{\kappa}} \right)^{-1} \xrightarrow{p} Id.$$

## K.2 Proof of Theorem 5.3.1

In order show consistency of  $\hat{\kappa}$ , we need to show that the combination of conditions (K1)-(K5) in Appendix K.1 and the conditions outlined directly in Theorem 5.3.1 imply conditions (F1)-(F4) from Appendix F.2 when  $\Psi_n(\kappa) = n^{-1} U^\dagger(\kappa, \hat{\theta})$ , as defined in (K1), and the data generating parameter is  $\kappa_0$ .

Condition (F2) is a direct result of (K1), and (F3) is a direct result of  $\Psi_n(\hat{\kappa}) \equiv 0$ .

To show (F1), it helps to define

$$\begin{aligned} g_j^{(1)}(t, \kappa, \theta) &= E_{\kappa, \theta}(h_j^d(t) | \bar{H}^d(t)), \\ g_j^{(2)}(t, \kappa, \theta) &= E_{\kappa, \theta}(\exp(\kappa^T h_j^d(t)) | \bar{H}^d(t)), \\ g_j^{(1)}(t, \kappa) &= g_j^{(1)}(t, \kappa, \theta_0), \\ g_j^{(2)}(t, \kappa) &= g_j^{(2)}(t, \kappa, \theta_0), \end{aligned}$$

and

$$S^{(1)}(t, \kappa, \theta) = \frac{1}{n} \sum_{j=1}^n \xi_j(t) g_j^{(1)}(t, \kappa, \theta) g_j^{(2)}(t, \kappa), \quad (\text{K.1})$$

$$S^{(0)}(t, \kappa, \theta) = \frac{1}{n} \sum_{j=1}^n \xi_j(t) g_j^{(2)}(t, \kappa, \theta), \quad (\text{K.2})$$

$$S^{(1)}(t, \kappa) = \frac{1}{n} \sum_{j=1}^n \xi_j(t) g_j^{(1)}(t, \kappa) g_j^{(2)}(t, \kappa),$$

$$S^{(0)}(t, \kappa) = \frac{1}{n} \sum_{j=1}^n \xi_j(t) g_j^{(2)}(t, \kappa).$$

Recalling the definition of  $\Psi_n^*(\kappa)$  in (K1), we may write

$$\Psi_n^*(\kappa) = \frac{1}{n} \sum_{i=1}^n \int_0^\tau \left[ g_i^{(1)}(t, \kappa) - \frac{S^{(1)}(t, \kappa)}{S^{(0)}(t, \kappa)} \right] dN_i^d(t),$$

and note that

$$\begin{aligned} & E_{\theta_0, \kappa_0}(\Psi_n^*(\kappa)) \tag{K.3} \\ &= E_{\theta_0, \kappa_0} \left( \frac{1}{n} \sum_{i=1}^n \int_0^\tau \left[ g_i^{(1)}(t, \kappa) - \frac{S^{(1)}(t, \kappa)}{S^{(0)}(t, \kappa)} \right] dN_i^d(t) \right) \\ &= E_{\theta_0, \kappa_0} \left( \frac{1}{n} \sum_{i=1}^n \int_0^\tau \left[ g_i^{(1)}(t, \kappa) - \frac{S^{(1)}(t, \kappa)}{S^{(0)}(t, \kappa)} \right] E_{\theta_0, \kappa_0} \left( dN_i^d(t) | \bar{H}^d(t) \right) \right) \\ &= E_{\theta_0, \kappa_0} \left( \frac{1}{n} \sum_{i=1}^n \int_0^\tau \left[ g_i^{(1)}(t, \kappa) - \frac{S^{(1)}(t, \kappa)}{S^{(0)}(t, \kappa)} \right] E_{\theta_0, \kappa_0} \left( E_{\theta_0, \kappa_0} \left( dN_i^d(t) | H^d(t) \right) | \bar{H}^d(t) \right) \right) \\ &\stackrel{(5.1)}{=} E_{\theta_0, \kappa_0} \left( \frac{1}{n} \sum_{i=1}^n \int_0^\tau \left[ g_i^{(1)}(t, \kappa) - \frac{S^{(1)}(t, \kappa)}{S^{(0)}(t, \kappa)} \right] \xi_i(t) g_i^{(2)}(t, \kappa_0) d\Lambda_0^d(t) \right) \\ &= (I) - (II), \end{aligned}$$

where

$$(I) = \int_0^\tau E_{\theta_0, \kappa_0} \left( \frac{1}{n} \sum_{i=1}^n g_i^{(1)}(t, \kappa) \xi_i(t) g_i^{(2)}(t, \kappa_0) \right) d\Lambda_0^d(t),$$

$$(II) = \int_0^\tau E_{\theta_0, \kappa_0} \left( \frac{S^{(1)}(t, \kappa)}{S^{(0)}(t, \kappa)} S^{(0)}(t, \kappa_0) \right) d\Lambda_0^d(t).$$

At  $\kappa = \kappa_0$ , we have  $(I) = (II)$  for all  $n$ ; hence,  $E_{\theta_0, \kappa_0}(\Psi_n^*(\kappa_0)) = 0$ . In view of the structure of (K.3), we further see that  $\Psi(\kappa)$  in (K1) is given by the following expression:

$$\Psi(\kappa) = \int_0^\tau [\omega_1(t, \kappa) - \omega_2(t, \kappa)] d\Lambda_0^d(t), \tag{K.4}$$

where

$$\begin{aligned}\omega_1(t, \kappa) &= E_{\theta_0, \kappa_0} \left( g_1^{(1)}(t, \kappa) \xi_1(t) g_1^{(2)}(t, \kappa_0) \right), \\ \omega_2(t, \kappa) &= \lim_{n \rightarrow \infty} E_{\theta_0, \kappa_0} \left( \frac{S^{(1)}(t, \kappa)}{S^{(0)}(t, \kappa)} S^{(0)}(t, \kappa_0) \right).\end{aligned}$$

Condition (F1) is now clear, and it only remains to show (F4). By the definitions in (K1) and a triangle inequality, the left hand side of (F4) in our current setting is bounded by

$$\sup_{\kappa \in \Pi} \|\Psi_n(\kappa) - \Psi_n^*(\kappa)\| + \sup_{\kappa \in \Pi} \|\Psi_n^*(\kappa) - \Psi(\kappa)\|. \quad (\text{K.5})$$

The first term in (K.5) converges in probability to zero by (K2), (K3), and the consistency of  $\hat{\theta}$ . The second term in (K.5) is more complicated. By (K3), the ratio of  $S^{(1)}(t, \kappa)$  and  $S^{(0)}(t, \kappa)$  converges uniformly to some vector valued function, say,  $s^{1,0}(t, \kappa)$ . By another triangle inequality, the second term in (K.5) is bounded by

$$\sup_{\kappa \in \Pi} \|\Psi_n^*(\kappa) - \Psi_n^{**}(\kappa)\| + \sup_{\kappa \in \Pi} \|\Psi_n^{**}(\kappa) - \Psi(\kappa)\|, \quad (\text{K.6})$$

where

$$\Psi_n^{**}(\kappa) = \frac{1}{n} \sum_{i=1}^n \int_0^\tau \left[ g_i^{(1)}(t, \kappa) - s^{1,0}(t, \kappa) \right] dN_i^d(t).$$

But now the first term in (K.6) converges in probability to zero by (G4) and (K3), and the second term converges in probability to zero by (K2) and (K3), which imply the conditions for Nishiyama (2009, Theorem 3.1 (ii)). Therefore (F4) is satisfied and this completes the proof of consistency.

To show asymptotic normality of  $\hat{\kappa}$  we need to show that the combination of conditions (K1)-(K5) in Appendix K.1 and the conditions outlined directly in Theorem 5.3.1 imply conditions (F1)-(F4) from Appendix F.2 when  $\Psi_n(\kappa) = n^{-1} U^\dagger(\kappa, \hat{\theta})$ , as defined in (K1), and the data generating parameter is  $\kappa_0$ .

Condition (F5) follows directly from the consistency of  $\hat{\kappa}$  and the fact that  $U^\dagger(\hat{\kappa}, \hat{\theta}) \equiv 0$ , and in our current setting, (F8) is a direct consequence of (K3), and

(F9) is a direct consequence of (K5). It remains to show (F6) and (F7), and also to estimate the asymptotic variance of  $\hat{\kappa}$ . To these ends, a sequence of Taylor series expansions will be helpful. First expand (4.9) around  $\theta_0$ :

$$0 = \tilde{U}^c(\hat{\theta}) = \tilde{U}^c(\theta_0) + \left. \frac{\partial \tilde{U}^c(\theta)}{\partial \theta} \right|_{\theta^*} (\hat{\theta} - \theta_0),$$

where  $\theta^*$  is on the line segment between  $\hat{\theta}$  and  $\theta_0$ . This yields

$$\hat{\theta} - \theta_0 = \left( - \left. \frac{\partial \tilde{U}^c(\theta)}{\partial \theta} \right|_{\theta^*} \right)^{-1} \tilde{U}^c(\theta_0).$$

Then expand (5.4) around  $(\kappa_0, \theta_0)$ :

$$U^\dagger(\kappa_0, \hat{\theta}) = U^\dagger(\kappa_0, \theta_0) + \left. \frac{\partial U^\dagger(\kappa_0, \theta)}{\partial \theta} \right|_{\theta^\circ} (\hat{\theta} - \theta_0),$$

where  $\theta^\circ$  is on the line segment between  $\hat{\theta}$  and  $\theta_0$ , which leads to

$$\frac{1}{\sqrt{n}} U^\dagger(\kappa_0, \hat{\theta}) = \frac{1}{\sqrt{n}} U^\dagger(\kappa_0, \theta_0) - \frac{1}{n} \left. \frac{\partial U^\dagger(\kappa_0, \theta)}{\partial \theta} \right|_{\theta^\circ} \left( \left. \frac{1}{n} \frac{\partial \tilde{U}^c(\theta)}{\partial \theta} \right|_{\theta^*} \right)^{-1} \frac{1}{\sqrt{n}} \tilde{U}^c(\theta_0). \quad (\text{K.7})$$

For convenience, define

$$D_7 = \lim_{n \rightarrow \infty} E \left[ - \left. \frac{1}{n} \frac{\partial U^\dagger(\kappa_0, \theta)}{\partial \theta} \right|_{\theta_0} \right],$$

and note that by (K4),

$$\hat{D}_7 \equiv - \left. \frac{1}{n} \frac{\partial U^\dagger(\kappa_0, \theta)}{\partial \theta} \right|_{\theta^\circ} \quad \text{and} \quad \hat{D}_5 \equiv - \left. \frac{1}{n} \frac{\partial \tilde{U}^c(\theta)}{\partial \theta} \right|_{\theta^*}$$

converge at rate  $n^{1/2}$  to  $D_7$  and  $D_5$  respectively. The reader is reminded here that

$$D_5 = \left. \frac{\partial}{\partial \theta} E_{\theta_0}(\tilde{U}_1^c(\theta)) \right|_{\theta_0}.$$

Defining

$$\bar{g}_j^{(1)}(t, \kappa, \theta) = \frac{S^{(1)}(t, \kappa, \theta)}{S^{(0)}(t, \kappa, \theta)},$$

we get

$$\begin{aligned} U^\dagger(\kappa_0, \theta_0) &= \sum_{i=1}^n \int_0^\tau \left( g_i^{(1)}(t, \kappa_0, \theta_0) - \bar{g}_i^{(1)}(t, \kappa_0, \theta_0) \right) dN_i^d(t) \\ &= \sum_{i=1}^n \int_0^\tau \left( g_i^{(1)}(t, \kappa_0, \theta_0) - \bar{g}_i^{(1)}(t, \kappa_0, \theta_0) \right) dM_i^d(t), \end{aligned}$$

where  $M_i^d(t) = N_i^d(t) - \int_0^t \xi_i(s) E_{\theta_0, \kappa_0}(\exp(\kappa_0^T h_i^d(s)) | \bar{H}^d(t)) d\Lambda_0^d(s)$  is a mean zero random variable. This follows from the unbiasedness calculation in (K.3).

Defining  $\tilde{U}^c(\theta) = \sum_{i=1}^n \int_0^\tau \tilde{\psi}_i(t, \theta) dN_i(t)$ , we see from (K.7) and the consistency of  $\hat{\theta}$  that  $n^{1/2} \Psi_n(\kappa_0) = n^{-1/2} U^\dagger(\kappa_0, \hat{\theta})$  is asymptotically equivalent to

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \int_0^\tau \left( g_i^{(1)}(t, \kappa_0, \theta_0) - g_0^{(1)}(t, \kappa_0, \theta_0) \right) dM_i^d(t) - D_7 D_5^{-1} \tilde{\psi}_i(t, \theta_0) dN_i(t), \quad (\text{K.8})$$

where  $g_0^{(1)}(t, \kappa_0, \theta_0)$  is the limit of  $\bar{g}_i^{(1)}(t, \kappa_0, \theta_0)$ . This asymptotic equivalence requires some remainder terms to converge to zero in probability, but this follows from (K4).

It is now evident from (G2) that (K.8) is a sum of i.i.d. mean zero random variables, and so by the central limit theorem, (F6) will be satisfied with

$$Z \sim N \left( 0, E \left[ \int_0^\tau \left( g_1^{(1)}(t, \kappa_0, \theta_0) - g_0^{(1)}(t, \kappa_0, \theta_0) \right) dM_1^d(t) - D_7 D_5^{-1} \tilde{\psi}_1(t, \theta_0) dN_1(t) \right]^{\otimes 2} \right),$$

where  $x^{\otimes 2} = xx^T$  for a vector  $x$ , as long as we can show that the summands of (K.8) have finite variance. But the variance of the portion with respect to  $dN_i(t)$  is finite by (G1)'s moment condition and (G4); the variance of the portion with respect to  $dM_i(t)$  is finite by (G1)'s left continuity condition, (G4), and (K3); and all the terms in the covariance of the two portions are finite by the Cauchy-Schwartz inequality and the finite variance results. Note that we also used (F1) to show the second part of the left hand side of (F6) is identically zero.

To show (F7), it helps to note that two of the terms in its numerator are zero. Hence it is sufficient to show:

$$\frac{n^{1/2} \|\Psi(\hat{\kappa}) - \Psi_n(\kappa_0)\|}{1 + n^{1/2} \|\hat{\kappa} - \kappa_0\|} \xrightarrow{p} 0. \quad (\text{K.9})$$

Some more Taylor expansions are now helpful. First,

$$\Psi(\hat{\kappa}) = \Psi(\kappa_0) + D_6(\hat{\kappa} - \kappa_0) + \frac{1}{2}(\hat{\kappa} - \kappa_0)^T \frac{\partial^2}{\partial \kappa \partial \kappa^T} \Psi(\kappa) \Big|_{\kappa_0} (\hat{\kappa} - \kappa_0), \quad (\text{K.10})$$

where the first term on the right hand side is zero and where  $\kappa^\circ$  is on the line segment between  $\hat{\kappa}$  and  $\kappa_0$ . Second,

$$0 = U^\dagger(\hat{\kappa}, \hat{\theta}) = U^\dagger(\kappa_0, \hat{\theta}) + \frac{\partial U^\dagger(\kappa, \hat{\theta})}{\partial \kappa} \Big|_{\kappa^*} (\hat{\kappa} - \kappa_0), \quad (\text{K.11})$$

where  $\kappa^*$  is on the line segment between  $\hat{\kappa}$  and  $\kappa_0$ . Rearranging (K.11) implies that:

$$\hat{\kappa} - \kappa_0 = \left( -\frac{\partial U^\dagger(\kappa, \hat{\theta})}{\partial \kappa} \Big|_{\kappa^*} \right)^{-1} U^\dagger(\kappa_0, \hat{\theta}). \quad (\text{K.12})$$

Combining (K.10) and (K.12), and using the triangle inequality, we see that the numerator of the left hand side of (K.9) is bounded by the sum of:

$$\left\| n^{-1/2} \left[ \left( -D_6 \left( -\frac{\partial}{\partial \kappa} \Psi_n(\kappa) \Big|_{\kappa^*} \right)^{-1} - Id \right) U^\dagger(\kappa_0, \hat{\theta}) \right] \right\| \quad (\text{K.13})$$

and

$$\left\| \frac{1}{2} n^{1/2} (\hat{\kappa} - \kappa_0)^T \frac{\partial^2}{\partial \kappa \partial \kappa^T} \Psi(\kappa) \Big|_{\kappa^\circ} (\hat{\kappa} - \kappa_0) \right\|. \quad (\text{K.14})$$

But (K.13) goes in probability to zero by combining (K5) and (F6), and by using the boundedness conditions on the second derivatives in (K3) and the consistency of  $\hat{\kappa}$ , (K.14) is bounded by a term that is  $o_P(1)n^{1/2}\|\hat{\kappa} - \kappa_0\|$ , where  $o_P(1)$  denotes convergence to zero in probability. This demonstrates (K.9), which implies (F7), and so the proof of asymptotic normality is complete.

Finally, using (F6) and the weak convergence result in Appendix F, note that the asymptotic variance of  $n^{1/2}(\hat{\kappa} - \kappa_0)$  is:

$$D_6^{-1} \left( E \left[ \int_0^\tau \left( g_1^{(1)}(t, \kappa_0, \theta_0) - g_0^{(1)}(t, \kappa_0, \theta_0) \right) dM_1^d(t) - D_7 D_5^{-1} \tilde{\psi}_1(t, \theta_0) dN_1(t) \right]^{\otimes 2} \right) (D_6^{-1})^T. \quad (\text{K.15})$$

An estimate of the variance is obtained by replacing the unknown quantities in (K.15) with their corresponding sample quantities.

## BIBLIOGRAPHY

Aalen, O. and Husebye, E. "Statistical Analysis of Repeated Events Forming Renewal Processes." *Statistics in Medicine*, 10:1227–1240 (1991).

Aalen, O. O., Fosen, J., Weedon-Fekjær, H., Borgan, Ø., and Husebye, E. "Dynamic analysis of multivariate failure time data." *Biometrics*, 60(3):764–773 (2004).

URL <http://dx.doi.org/10.1111/j.0006-341X.2004.00227.x>

Aalen, O. O. and Gunnes, N. "A dynamic approach for reconstructing missing longitudinal data using the linear increments model." *Biostatistics*, 11(3):453–72 (2010).

URL <http://dx.doi.org/10.1093/biostatistics/kxq014>

Aït-Sahalia, Y. "Estimating continuous-time models with discretely sampled data." Invited Lecture, World Congress of the Econometric Society (2005).

Aït-Sahalia, Y. and Mykland, P. A. "The effects of random and discrete sampling when estimating continuous-time diffusions." *Econometrica*, 71(2):483–549 (2003).

URL <http://dx.doi.org/10.1111/1468-0262.t01-1-00416>

Andersen, P. K. and Gill, R. D. "Cox's regression model for counting processes: a large sample study." *Ann. Statist.*, 10(4):1100–1120 (1982).

Andersen, P. K. and Liestøl, K. "Attenuation caused by infrequently updated covariates in survival analysis." *Biostatistics*, 4(4):633–49 (2003).

Bang, H. and Robins, J. M. "Doubly robust estimation in missing data and causal inference models." *Biometrics*, 61(4):962–972 (2005).

URL <http://dx.doi.org/10.1111/j.1541-0420.2005.00377.x>



- Bibby, B. M., Jacobsen, M., and Sørensen, M. "Estimating functions for discretely sampled diffusion-type models." (2004).
- Bůžková, P. and Lumley, T. "Semiparametric modeling of repeated measurements under outcome-dependent follow-up." *Statistics in Medicine*, 28:987–1003 (2009).
- Chang, S.-H. "Estimating marginal effects in accelerated failure time models for serial sojourn times among repeated events." *Lifetime Data Anal.*, 10(2):175–190 (2004).  
URL <http://dx.doi.org/10.1023/B:LIDA.0000030202.20842.c9>
- Chang, S.-H. and Wang, M.-C. "Conditional regression analysis for recurrence time data." *J. Amer. Statist. Assoc.*, 94(448):1221–1230 (1999).
- Chen, Y. and Wang, M. "Semiparametric regression analysis on longitudinal pattern of recurrent gap times." *Biostatistics*, 5:277–290 (2004).
- Clement, D. Y. and Strawderman, R. L. "Conditional GEE for recurrent event gap times." *Biostatistics*, 10(3):451–467 (2009).  
URL <http://dx.doi.org/10.1093/biostatistics/kxp004>
- Cook, R. and Lawless, J. *The Statistical Analysis of Recurrent Events*. New York: Springer-Verlag (2007).
- Cox, D. R. *Renewal theory*. London: Methuen & Co. Ltd. (1962).
- . "Regression models and life-tables." *J. Roy. Statist. Soc. Ser. B*, 34:187–220 (1972). With discussion by F. Downton, Richard Peto, D. J. Bartholomew, D. V. Lindley, P. W. Glassborow, D. E. Barton, Susannah Howard, B. Benjamin, John J. Gart, L. D. Meshalkin, A. R. Kagan, M. Zelen, R. E. Barlow, Jack Kalbfleisch, R. L. Prentice and Norman Breslow, and a reply by D. R. Cox.

- Dafni, U. G. and Tsiatis, A. A. "Evaluating surrogate markers of clinical outcome when measured with error." *Biometrics*, 54(4):1445–1462 (1998).
- Dempster, A. P., Laird, N. M., and Rubin, D. B. "Maximum likelihood from incomplete data via the EM algorithm." *J. Roy. Statist. Soc. Ser. B*, 39(1):1–38 (1977). With discussion.
- Diggle, P. J., Heagerty, P., Liang, K.-Y., and Zeger, S. *Analysis of Longitudinal Data*. New York: Oxford University Press, 2nd edition (2002).
- Dubin, J. A. and Müller, H.-G. "Dynamical correlation for multivariate longitudinal data." *J. Amer. Statist. Assoc.*, 100(471):872–881 (2005).  
URL <http://dx.doi.org/10.1198/016214504000001989>
- Duchateau, L., Janssen, P., Kezic, I., and Fortpied, C. "Evolution of recurrent asthma event rate over time in frailty models." *J. Roy. Statist. Soc. Ser. C*, 52(3):355–363 (2003).  
URL <http://dx.doi.org/10.1111/1467-9876.00409>
- Elashoff, M. and Ryan, L. "An EM algorithm for estimating equations." *J. Comput. Graph. Statist.*, 13(1):48–65 (2004).  
URL <http://dx.doi.org/10.1198/1061860043092>
- Fieuws, S. and Verbeke, G. "Pairwise fitting of mixed models for the joint modeling of multivariate longitudinal profiles." *Biometrics*, 62(2):424–431 (2006).  
URL <http://dx.doi.org/10.1111/j.1541-0420.2006.00507.x>
- Fine, J. P., Yan, J., and Kosorok, M. R. "Temporal process regression." *Biometrika*, 91(3):683–703 (2004).  
URL <http://dx.doi.org/10.1093/biomet/91.3.683>

- Fleming, T. R. and Harrington, D. P. *Counting processes and survival analysis*. New York: Wiley (1991).
- Foutz, R. V. "On the unique consistent solution to the likelihood equations." *J. Amer. Statist. Assoc.*, 72(357):147–148 (1977).
- Gerds, T. A. and Schumacher, M. "Consistent estimation of the expected Brier score in general survival models with right-censored event times." *Biom. J.*, 48(6):1029–1040 (2006).  
URL <http://dx.doi.org/10.1002/bimj.200610301>
- . "Efron-type measures of prediction error for survival analysis." *Biometrics*, 63(4):1283–1287, 1316 (2007).
- Gill, R. D. and Grünwald, P. D. "An algorithmic and a geometric characterization of coarsening at random." *Ann. Statist.*, 36(5):2409–2422 (2008).  
URL <http://dx.doi.org/10.1214/07-AOS532>
- Gill, R. D., Laan, M. J. V. D., and Robins, J. M. "Coarsening At Random: Characterizations, Conjectures, Counter-Examples." In *Proceedings of the First Seattle Symposium on Survival Analysis*, 255–294. Springer Verlag (1996).
- Heitjan, D. F. and Rubin, D. B. "Ignorability and coarse data." *Ann. Statist.*, 19(4):2244–2253 (1991).  
URL <http://dx.doi.org/10.1214/aos/1176348396>
- Hogan, J., Roy, J., and Korkontzelou, C. "Handling drop-out in longitudinal studies." *Statistics in medicine*, 23(9):1455–1497 (2004).  
URL <http://dx.doi.org/10.1002/sim.1728>
- Huang, Y. "Censored regression with the multistate accelerated sojourn times

- model." *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 64(1):17–29 (2002).  
 URL <http://dx.doi.org/10.1111/1467-9868.00322>
- Huang, Y. and Chen, Y. Q. "Marginal regression of gaps between recurrent events." *Lifetime Data Anal.*, 9(3):293–303 (2003).  
 URL <http://dx.doi.org/10.1023/A:1025892922453>
- Kalbfleisch, J. D. and Prentice, R. L. *The Statistical Analysis of Failure Time Data*. Hoboken, New Jersey: Wiley Interscience, 2nd edition (2002).
- Kaysen, G. A., Dubin, J. A., Müller, H. G., Rosales, L. M., and Levin, N. W. "The acute-phase response varies with time and predicts serum albumin levels in hemodialysis patients. The HEMO Study Group." *Kidney international*, 58(1):346–352 (2000).  
 URL <http://dx.doi.org/10.1046/j.1523-1755.2000.00172.x>
- Kessler, M. "Simple and explicit estimating functions for a discretely observed diffusion process." *Scand. J. Statist.*, 27(1):65–82 (2000).  
 URL <http://dx.doi.org/10.1111/1467-9469.00179>
- Kessler, M. and Sørensen, M. "Estimating equations based on eigenfunctions for a discretely observed diffusion process." *Bernoulli*, 5(2):299–314 (1999).  
 URL <http://dx.doi.org/10.2307/3318437>
- Kosorok, M. R. *Introduction to Empirical Processes and Semiparametric Inference*. New York: Springer (2008).
- Kotz, S. and Shanbhag, D. N. "Some new approaches to probability distributions." *Adv. in Appl. Probab.*, 12(4):903–921 (1980).  
 URL <http://dx.doi.org/10.2307/1426748>

Liang, K. Y. and Zeger, S. L. "Longitudinal data analysis using generalized linear models." *Biometrika*, 73(1):13–22 (1986).

URL <http://dx.doi.org/10.1093/biomet/73.1.13>

Liang, Y., Lu, W., and Ying, Z. "Joint modeling and analysis of longitudinal data with informative observation times." *Biometrics* (2009).

Lin, D. Y., Sun, W., and Ying, Z. "Nonparametric estimation of the gap time distributions for serial events with censored data." *Biometrika*, 86(1):59–70 (1999).

URL <http://dx.doi.org/10.1093/biomet/86.1.59>

Lin, D. Y. and Ying, Z. "Semiparametric and nonparametric regression analysis of longitudinal data." *J. Amer. Statist. Assoc.*, 96(453):103–126 (2001). With comments and a rejoinder by the authors.

URL <http://dx.doi.org/10.1198/016214501750333018>

Lin, H., Scharfstein, D. O., and Rosenheck, R. A. "Analysis of longitudinal data with irregular, outcome-dependent follow-up." *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 66(3):791–813 (2004).

URL <http://dx.doi.org/10.1111/j.1467-9868.2004.b5543.x>

Lin, H., Turnbull, B. W., McCulloch, C. E., and Slate, E. H. "Latent class models for joint analysis of longitudinal biomarker and event process data: application to longitudinal prostate-specific antigen readings and prostate cancer." *J. Amer. Statist. Assoc.*, 97(457):53–65 (2002).

URL <http://dx.doi.org/10.1198/016214502753479220>

Lipsitz, S. R., Fitzmaurice, G. M., Ibrahim, J. G., Gelber, R., and Lipshultz, S. "Parameter estimation in longitudinal studies with outcome-dependent follow-

- up." *Biometrics*, 58(3):621–630 (2002).  
 URL <http://dx.doi.org/10.1111/j.0006-341X.2002.00621.x>
- Lipsitz, S. R., Ibrahim, J. G., and Zhao, L. P. "A weighted estimating equation for missing covariate data with properties similar to maximum likelihood." *J. Amer. Statist. Assoc.*, 94(448):1147–1160 (1999).
- Liu, L., Conaway, M. R., Knaus, W. A., and Bergin, J. D. "A random effects four-part model, with application to correlated medical costs." *Computational Statistics & Data Analysis*, 52(9):4458–4473 (2008a).
- Liu, L., Huang, X., and O’Quigley, J. "Analysis of longitudinal data in the presence of informative observational times and a dependent terminal event, with application to medical cost data." *Biometrics*, 64(3):950–958 (2008b).  
 URL <http://dx.doi.org/10.1111/j.1541-0420.2007.00954.x>
- Molenberghs, G. and Verbeke, G. *Models for Discrete Longitudinal Data*. New York: Springer (2005).
- Murphy, S. and Li, B. "Projected partial likelihood and its application to longitudinal data." *Biometrika*, 82(2):399–406 (1995).  
 URL <http://dx.doi.org/10.1093/biomet/82.2.399>
- Murphy, S. A., Bentley, G. R., and O’Hanesian, M. A. "An analysis for menstrual data with time-varying covariates." *Statistics in Medicine*, 14:1843–1857 (1995).
- Newey, W. K. "Uniform convergence in probability and stochastic equicontinuity." *Econometrica*, 59(4):1161–1167 (1991).  
 URL <http://dx.doi.org/10.2307/2938179>
- Nishiyama, Y. "Asymptotic theory of semiparametric Z-estimators for stochastic processes with applications to ergodic diffusions and time series." *Ann.*

- Statist.*, 37(6A):3555–3579 (2009).  
 URL <http://dx.doi.org/10.1214/09-AOS693>
- Oakes, D. and Cui, L. “On semiparametric inference for modulated renewal processes.” *Biometrika*, 81(1):83–90 (1994).  
 URL <http://dx.doi.org/10.2307/2337052>
- Oakes, D. and Dasu, T. “A note on residual life.” *Biometrika*, 77(2):409–410 (1990).  
 URL <http://dx.doi.org/10.1093/biomet/77.2.409>
- Peña, E. A., Strawderman, R. L., and Hollander, M. “Nonparametric estimation with recurrent event data.” *J. Amer. Statist. Assoc.*, 96(456):1299–1315 (2001).  
 URL <http://dx.doi.org/10.1198/016214501753381922>
- Pepe, M. S. and Anderson, G. L. “A cautionary note on inference for marginal regression models with longitudinal data and general correlated response data.” *Communications in Statistics, Part B Simulation and Computation*, 23:939–951 (1994).
- Prentice, R. L. “Correlated binary regression with covariates specific to each binary observation.” *Biometrics*, 44(4):1033–1048 (1988).  
 URL <http://www.hubmed.org/display.cgi?uids=3233244>
- Prentice, R. L., Williams, B. J., and Peterson, A. V. “On the regression analysis of multivariate failure time data.” *Biometrika*, 68(2):373–379 (1981).  
 URL <http://dx.doi.org/10.1093/biomet/68.2.373>
- Prentice, R. L. and Zhao, L. P. “Estimating equations for parameters in means and covariances of multivariate discrete and continuous responses.” *Biomet-*

*rics*, 47(3):825–839 (1991).

URL <http://dx.doi.org/10.2307/2532642>

Rizopoulos, D., Verbeke, G., and Lesaffre, E. “Fully exponential Laplace approximations for the joint modelling of survival and longitudinal data.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71:637–654(18) (2009).

Robins, J. M., Rotnitzky, A., and Zhao, L. P. “Analysis of semiparametric regression models for repeated outcomes in the presence of missing data.” *J. Amer. Statist. Assoc.*, 90(429):106–121 (1995).

URL <http://www.jstor.org/stable/2291134>

Rotnitzky, A., Robins, J. M., and Scharfstein, D. O. “Semiparametric regression for repeated outcomes with nonignorable nonresponse.” *J. Amer. Statist. Assoc.*, 93(444):1321–1339 (1998).

Rubin, D. B. “Inference and missing data.” *Biometrika*, 63(3):581–592 (1976).  
With comments by R. J. A. Little and a reply by the author.

—. “Multiple imputation after 18+ years.” *J. Amer. Statist. Assoc.*, 91(434):473–489 (1996).

Scharfstein, D. O. and Robins, J. M. “Estimation of the failure time distribution in the presence of informative censoring.” *Biometrika*, 89(3):617–634 (2002).

URL <http://dx.doi.org/10.1093/biomet/89.3.617>

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. “Adjusting for nonignorable drop-out using semiparametric nonresponse models.” *J. Amer. Statist. Assoc.*, 94(448):1096–1146 (1999). With comments and a rejoinder by the authors.



Schoop, R. "Predictive accuracy of failure time models with longitudinal covariates." Ph.D. thesis, Albert-Ludwigs-Universität Freiburg (2008).

Schoop, R., Graf, E., and Schumacher, M. "Quantifying the predictive performance of prognostic models for censored survival data with time-dependent covariates." *Biometrics*, 64(2):603–610, 670 (2008).

URL <http://dx.doi.org/10.1111/j.1541-0420.2007.00889.x>

Self, S. and Pawitan, Y. "Modeling a marker of disease progression and onset of disease." In Jewell, N., Dietz, K., and Farewell, V. (eds.), *in AIDS Epidemiology: Methodological Issues*. Birkhauser, Boston (1992).

Strawderman, R. L. "The accelerated gap times model." *Biometrika*, 92(3):647–666 (2005).

URL <http://dx.doi.org/10.1093/biomet/92.3.647>

—. "A regression model for dependent gap times." *Int. J. Biostat.*, 2:Art. 1, 34 pp. (electronic) (2006).

Sun, J., Park, D.-H., Sun, L., and Zhao, X. "Semiparametric regression analysis of longitudinal data with informative observation times." *J. Amer. Statist. Assoc.*, 100(471):882–889 (2005).

URL <http://dx.doi.org/10.1198/016214505000000060>

Sun, J., Sun, L., and Liu, D. "Regression analysis of longitudinal data in the presence of informative observation and censoring times." *J. Amer. Statist. Assoc.*, 102(480):1397–1406 (2007).

URL <http://dx.doi.org/10.1198/016214507000000851>

Sun, L. and Tong, X. "Analyzing longitudinal data with informative observation

- times under biased sampling." *Statistics and Probability Letters*, 79:1162–1168 (2009).
- Tsiatis, A., DeGruttola, V., and Wulfsohn, M. "Modeling the Relationship of Survival to Longitudinal Data Measured with Error. Applications to Survival and CD4 Counts in Patients with AIDS." *J. Amer. Statist. Assoc.*, 90:27–37, (1995).
- Tsiatis, A. A. and Davidian, M. "A semiparametric estimator for the proportional hazards model with longitudinal covariates measured with error." *Biometrika*, 88(2):447–458 (2001).  
URL <http://dx.doi.org/10.1093/biomet/88.2.447>
- . "Joint modeling of longitudinal and time-to-event data: an overview." *Statist. Sinica*, 14(3):809–834 (2004).
- van der Vaart, A. *Asymptotic Statistics*. New York: Cambridge University Press (1998).
- Van Houwelingen, H. C. "Dynamic prediction by landmarking in event history analysis." *Scand. J. Statist.*, 34(1):70–85 (2007).  
URL <http://dx.doi.org/10.1111/j.1467-9469.2006.00529.x>
- Wang, Y. and Taylor, J. M. G. "Jointly modeling longitudinal and event time data with application to acquired immunodeficiency syndrome." *J. Amer. Statist. Assoc.*, 96(455):895–905 (2001).  
URL <http://dx.doi.org/10.1198/016214501753208591>
- Wulfsohn, M. S. and Tsiatis, A. A. "A joint model for survival and longitudinal data measured with error." *Biometrics*, 53(1):330–339 (1997).  
URL <http://dx.doi.org/10.2307/2533118>

- Xu, J. and Zeger, S. L. "Joint analysis of longitudinal data comprising repeated measures and times to events." *J. Roy. Statist. Soc. Ser. C*, 50(3):375–387 (2001).  
URL <http://dx.doi.org/10.1111/1467-9876.00241>
- Yuan, K.-H. and Jennrich, R. I. "Asymptotics of estimating equations under natural conditions." *J. Multivariate Anal.*, 65(2):245–260 (1998).  
URL <http://dx.doi.org/10.1006/jmva.1997.1731>
- Zhao, L. P. and Prentice, R. L. "Correlated binary regression using a quadratic exponential model." *Biometrika*, 77(3):642–648 (1990).  
URL <http://dx.doi.org/10.1093/biomet/77.3.642>
- Zheng, Y. and Heagerty, P. J. "Partly conditional survival models for longitudinal data." *Biometrics*, 61(2):379–391 (2005).  
URL <http://dx.doi.org/10.1111/j.1541-0420.2005.00323.x>
- Zorn, C. J. W. "Generalized Estimating Equation Models for Correlated Data: A Review with Applications." *American Journal of Political Science*, 45(2):470–490 (2001).  
URL <http://dx.doi.org/10.2307/2669353>